Characterisation of the *WAK / WAKL* gene family, with a focus on members in *Persea americana* implicated in defense against *Phytophthora cinnamomi*

By

Aaron Thomas Harvey
19079843

Submitted in fulfilment of the requirements for the degree

Magister Scientiae

In the Faculty of Natural and Agricultural Sciences
Department of Biochemistry, Genetics and Microbiology
University of Pretoria
South Africa

Supervisors: Dr Velushka Swart
Co-supervisor: Prof. Noëlani van den Berg

31 October 2024

Table of Contents:

Declaration of Originality iii

Ethics statement iv

Acknowledgements v

Summary vi

Research outputs vii

List of abbreviations viii

Standard Units of Measure and Symbols ix

Chapter 1 - Describing and characterising the WAK/WAKL gene family across plant species: A systematic review 1

- Introduction 2
- Identification of WAKs and WAKLs 3
  - Discovery of WAK/WAKLs and protein domain composition 3
  - Current approaches and inconsistency in the identification and classification of WAK/WAKLs 3
- *WAK* and *WAKL* composition of species recently characterised 5
- Gene comparisons 6
  - *WAK/WAKL* gene structure, placement and duplication methods 6
  - *Cis*-acting element comparison in *WAK/WAKL* promotor regions 8
- Expression data comparison 10
  - Tissue specificity and developmental stages 10
  - Abiotic and biotic responses 11
- Protein comparisons 11
  - General properties and subcellular localisation 11
  - Phylogenetic analyses for functional inference and conserved motif analyses 13
  - Impact of polymorphism on function 14
- Discussion 14
- References 16

Chapter 2 - *In silico* characterisation of the avocado *WAK/WAKL* gene family with a focus on genes involved in defence against *Phytophthora cinnamomi* 18

- Abstract 19
- Introduction 20
- Materials and methods 21
  - Putative *PaWAK/WAKL* identification 21
  - *PaWAK/WAKL* classification 22
  - PaWAK/WAKL protein sequence analyses 22

PaWAK/WAKL expression analysis during *P. cinnamomi* infection 23
Phylogenetic analysis of PaWAK/WAKLs 23
Promoter analysis of *PaWAK/WAKLs* 23
Structure and ligand binding analyses of candidate defence PaWAK/WAKLs 24
Results 25
The full repertoire of *PaWAK* and *PaWAKL* genes characterised 25
PaWAK/WAKL protein properties and conserved motifs 28
Expression analysis of *PaWAK/WAKLs* during *P. cinnamomi* infection 30
Phylogenetic analysis of PaWAK/WAKL, AtWAK/WAKL and defence-related WAK/WAKLs 32
Prediction of *cis*-acting elements for inferring involvement in cellular pathways 34
Structure and ligand binding analyses of candidate defence PaWAK/WAKLs 36
Discussion 39
In silico characterisation of the *PaWAK/WAKL* gene family 39
*PaWAK/WAKL*s implicated in avocado defence against *Phytophthora cinnamomi* 41
Candidate defence PaWAK/WAKL analyses show structural and ligand-binding differences 43
Conclusion 45
Supplementary Material 46
References 47

Declaration of Originality

I, Aaron Thomas Harvey, declare that the dissertation, which I hereby submit for the degree MSc Genetics at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Signature:

31 October 2024

Ethics statement

The author, whose name appears on the title page of this dissertation/thesis, has obtained, for the research described in this work, the applicable research ethics approval. Ethics approval number: NAS118/2023.

The author declares that he has observed the ethical standards required in terms of the University of Pretoria's Code of Ethics for Researchers and the Policy guidelines for responsible research.

Signature:

Aaron Harvey
31 October 2024

Acknowledgements

- I would like to thank my supervisor, Dr Velushka Swart, a million times over for being so supportive and encouraging. You have become my academic mother, and I feel your nurturing ways every time I walk into your office. A simple 'How are things going' can have us talking for hours, but I would not have it any other way. I attribute the scientist I am today to your guidance, patience and care. I will always be eternally grateful that fate put us together.
- I would like to thank Prof. Noëlani van den Berg for being by academic aunt. You are always there on the side lines making sure I do not fall. Your careful watch has made me feel safe and enjoy coming into the office/lab every day.
- Thank you to the rest of the Hans Merensky Chair in Avocado research senior team for your continued support. A special thanks to Dr Robert Backer for your continued bioinformatic support.
- I would not be here if it were not for the continued love and support of my family. Mom, Dad, Josh, Julia, James, Kim and Callan, I love you all and do not have the words to thank you for all that you have done and for the impact you have had in my life.
- Thank you, Michael, for the continuous laughs and for being my confidant. You helped me realise that there is more to life than work.
- Amper Dr Fick, you have been my rock. I cannot thank you enough for just being you.
- To my lab acquaintances, Aly, Kayla, Raven, JJ, Kat: Thank you for bringing the sass, dealing with my sass, and making coming into work more of a pleasure.
- To the rest of the 'Avo Gang', thank you all for the amazing sense of togetherness and community. Together we will accomplish amazing things.
- Thank you to the Hans Merensky Legacy Foundation for funding this academic journey.

Summary

Phytophthora cinnamomi is a devastating oomycete pathogen, which has a significant negative impact on global avocado (*Persea americana)* production. The wall associated kinases (WAKs) and wall associated kinase-likes (WAKLs) are pattern recognition receptors which detect fragmented pectin (oligogalacturonides formed during pathogen infection) and activate downstream plant defence responses. The identification and charaterisation of this gene family shows many inconsistencies which needed to be addressed before delving into the avocado *WAK/WAKL* gene family and describing which member are implicated in the defence response of *P. cinnamomi.* Thus, this dissertation aimed to critically evaluate the *in silico* characterization of the *WAK* and *WAKL* gene family as a whole and then in avocado, with a particular focus on the genes involved in defence against *P. cinnamomi*.

The study begins with a critical evaluation of current methodologies for identifying and classifying *WAK/WAKL* genes across various plant species, revealing inconsistencies in identification, classification and characterisation protocols that hinder broader gene family conclusions. The review provides a streamlined approach to help increase the standardisation in this gene family so that all genes defined as *WAK*s or *WAKL*s contain the same properties, such as a standardised set of predicted protein domains in the downstream proteins. The review also provides recommendations for characterising the family to allow a broader description to be concatenated for large-scale analyses in the future.

The core of the research is the comprehensive *in silico* characterization of the avocado *WAK/WAKL* (*PaWAK/WAKL*) gene family. The study describes the *PaWAK/WAKL* family composition, gene placement and structure, the protein properties and cellular localisation, phylogenetic relationships, their expression patterns during pathogen infection, and the potential impact of *cis*-acting elements in targeted regulation during defence. Protein 3D structures and the binding affinities of WAK/WAKLs and the damage-associated molecular pattern, oligogalacturonide, were predicted for defence-implicated proteins. This dissertation examines the *PaWAK/WAKL* on a genomic, transcriptomic, and proteomic level to identify differences between the partially-resistant Dusa® and susceptible R0.12 rootstocks (while

also assessing the partially-resistant Leola™) to predict how these factors, in combination, contribute to the increased defence efficiency against *P. cinnamomi*.

These findings highlight the significance of specific *PaWAK/WAKL* genes in avocado's defence against *P. cinnamomi* on a multi-omic level. This research not only enhances our understanding of this disease interaction but also serves as a foundation for developing molecular tools to screen for resistant rootstocks for commercial use.

Research outputs

Harvey A, van den Berg N and Swart V (2024). Describing and characterizing the *WAK/WAKL* gene family across plant species: a systematic review. *Front. Plant Sci*. 15:1467148. doi: 10.3389/fpls.2024.1467148

Harvey A., van den Berg N., Swart V. *In silico* characterisation of the avocado *WAK/WAKL* gene family with a focus on genes involved in defence against *Phytophthora cinnamomi*. *Submitted to Frontiers in Plant Science.*

Swart V., Fick A., Anbu S., Harvey A., Moodley S., van den Berg N. Unravelling avocado defence against *Phytophthora cinnamomi*. Poster presentation at Plant Pathology 2024 Annual meeting of the British Society for Plant Pathology; 2024 Sep 11-13, Oxford, United Kingdom.

Harvey A., van den Berg N., Swart V. *In silico* characterisation of the *WAK/WAKL* gene family in avocado and implication in defence against *Phytophthora cinnamomi*. Oral presentation at the 20th International Plant Protection Congress; 2024 Jul 1-5, Athens, Greece.

Harvey A., van den Berg N., Swart V. Genome-wide *in silico* characterisation of the avocado *WAK* and *WAKL* gene family. Oral presentation at the 53rd biennial congress of the Southern African Society for Plant Pathology; 2024 Jan 22-25, Golden Gate National Park, Free state, South Africa.

List of abbreviations

ABA – Abscisic acid

BA – Binding affinity

BLAST – Basic Local Alignment Search Tool

CDS – Coding sequence (Exons)

DAMP – Damage associated molecular pattern

Desolv – Desolvation energy

DNA - Deoxyribonucleic acid

EGF – Epidermal growth factor

EGF_CA - Epidermal growth factor calcium binding

ERE – Ethylene

GFP – Green fluorescent protein

GMO – Genetically modified organism

GWB – GUB_WAK_bind – homogalacturonan-binding region

Hbond – Hydrogen bonds

HMM – Hidden Markov model

JA – Jasmonic acid

K_a – Synonymous mutation rate

K_a/K_s – Nonsynonymous-to-synonymous ratio

K_d – Dissociation constant

K_i – Inhibition constant

K_s – Nonsynonymous mutation rate

LBn – Local BLAST for nucleotides

LBp – Local BLAST for proteins

Log2FC – Log2FoldChange

LTR – Low-temperature responsive

MeJA – Methyl Jasmonate

MEME – Multiple EM for Motif Elicitation

NCBI CDD – National Center for Biotechnology Information Conserved Domains Database

OG – Oligogalacturonide

ORF – Open reading frame

PDB file – Protein Data Bank file

PG – Polygalacturonase

pk – pkinase

PKc-like – Protein Kinase, cytoplasmic domain-like

RMSD – Root-mean-square deviation

RLK – Receptor-like kinase

RNA – Ribonucleic acid

ROS – Reactive oxygen species

RT-qPCR – Reverse transcriptase quantitative polymerase chain reaction

SA – Salicylic acid

SMART – Simple Modular Architecture Research Tool

SP – Signal peptide

STK – Serine/Threonine kinase

STKc_IPAK – Catalytic domain of the Serine/Threonine kinases, Interleukin-1 Receptor Associated Kinases and related STKs

TM – Transmembrane domain

UPGMA – Unweighted Pair-Group Method with Arithmetic Averaging

UTR – Untranslated regions

vdW – van der Waals forces

VIGS - Virus-induced gene silencing

WAK – Wall-associated kinase

WAKa – WAK associated

WAKL – Wall-associated kinase-like

Standard Units of Measure and Symbols

Å – Armstrongs

Aa – amino acids

Bp – base pairs

Kb – Kilobases

kcal/mol - kilocalorie per mole

kDa – Kilodaltons

Mb – Mega bases

Chapter 1 - Describing and characterising the WAK/WAKL gene family across plant species: A systematic review

Published in Frontiers in Plant Science – Plant Bioinformatics

Volume 15 - 2024 | doi: 10.3389/fpls.2024.1467148

frontiers | Frontiers in Plant Science

TYPE Review
PUBLISHED 12 November 2024
DOI 10.3389/fpls.2024.1467148

OPEN ACCESS

EDITED BY
Deepak Singla,
Punjab Agricultural University, India

REVIEWED BY
Chuan Xia,
Chinese Academy of Agricultural Sciences (CAAS), China
Pawan Kumar,
Chaudhary Charan Singh Haryana Agricultural University, India
Arvind K Yadav,
Indian Council of Agricultural Research (ICAR), India

*CORRESPONDENCE
Velushka Swart
velushka.swart@up.ac.za

RECEIVED 19 July 2024
ACCEPTED 22 October 2024
PUBLISHED 12 November 2024

CITATION
Harvey A, van den Berg N and Swart V (2024) Describing and characterizing the *WAK/WAKL* gene family across plant species: a systematic review.
Front. Plant Sci. 15:1467148.
doi: 10.3389/fpls.2024.1467148

Describing and characterizing the *WAK/WAKL* gene family across plant species: a systematic review

Aaron Harvey, Noëlani van den Berg and Velushka Swart*

Hans Merensky Chair in Avocado Research, Forestry & Agricultural Biotechnology Institute (FABI), Department of Biochemistry, Genetics and Microbiology (BGM), University of Pretoria, Pretoria, South Africa

Wall-associated kinases (WAKs) and WAK-likes (WAKLs) are transmembrane pectin receptors which have seen rising interest in recent years due to their roles in stress responses and developmental pathways. Consequently, the genes encoding these proteins are continuously identified, described and characterised across a wide variety of plant species. The primary goal of characterizing these genes is to classify, describe and infer cellular function, mostly through *in silico* methods. However, inconsistencies across characterizations have led to discrepancies in WAK/WAKL definitions resulting in sequences being classified as a WAK in one study but as a WAKL or not identified in another. The methods of characterization range widely with different combinations of analyses being conducted, to similar analyses but with varying inputs and parameters which are impacting the outputs. This review collates current knowledge about *WAK/WAKL* genes and the recent characterizations of this family and suggests a more robust strategy for increased consistency among the different gene members, as well as the characterizations thereof.

KEYWORDS

wall-associated kinase, wall-associated kinase-like, gene identification and classification, expression data, *cis*-acting elements

1 Introduction

Wall-associated kinase (WAK) proteins are transmembrane pectin receptors that constitute one of the 15 receptor-like kinase (RLK) subfamilies and serve multiple functions within plant cells (Stephens et al., 2022). The first WAK protein (AtWAK1) was described in *Arabidopsis thaliana* using immunohistochemistry to confirm its association with the cell wall, and protease experiments to confirm the cytoplasmic kinase domain, thereby defining a protein capable of facilitating signaling between the cell wall and the cytoplasm (He et al., 1996). Subsequently, WAK-like (WAKL) proteins were also described in *A. thaliana* (Verica and He, 2002).

WAKs bind pectin in multiple forms, such as native pectin to influence cell expansion during plant development, and oligogalacturonides (fragmented pectin) that act as damage-associated molecular patterns (DAMPs) during abiotic and biotic stresses (Kohorn and Kohorn, 2012; Kohorn, 2015). Some WAK/WAKLs have been shown to confer resistance to hemibiotrophic and necrotrophic fungi through a range of mechanisms including pathogen- or host-derived elicitor detection and cell wall restructuring (Stephens et al., 2022).

The original *WAK/WAKL* family identified in *A. thaliana* consisted of five *AtWAKs* and 21 *AtWAKLs*. The *WAK/WAKL* gene families have since been characterised in many plant species, including *Solanum tuberosum* (potato), *Gossypium hirsutum* (cotton) and *Triticum aestivum* (bread wheat) which contained 29, 99 and 320 *WAK* and/or *WAKL* genes, respectively (Xia et al., 2022; Yu et al., 2022; Zhang et al., 2021).

Over the past two decades the *WAK/WAKL* gene family has been a subject of interest due to its involvement in plant defense and development (Stephens et al., 2022). The gene family has been bioinformatically and functionally characterised across several plant species to identify candidate genes involved in the successful defense response, highlighting their potential for improving crop disease resistance (Stephens et al., 2022). Recent characterizations of this gene family typically include the number of *WAK/WAKLs* identified within a species, predicted protein properties, phylogenetic analysis, gene structure visualization, and chromosome placement. The characterization studies also provide information on expression patterns in different tissues, under stress conditions, phytohormone introduction, the presence of conserved protein motifs and *cis*-acting elements in gene promoter regions, selection pressure predictions, subcellular localization and duplication predictions.

This review will synthesize and compare the *WAK/WAKL* gene families across plant species, the methodology used to identify, classify and characterize these genes, and propose improvements to increase consistency and comparability between different plant species and studies.

2 Identification of WAKs and WAKLs

2.1 Discovery of WAK/WAKLs and protein domain composition

AtWAK1 was first described as a transmembrane cell-wall associated protein with a cytoplasmic kinase domain (the serine/threonine protein kinase domain, pkinase) and extracellular epidermal growth factor (EGF)-containing regions (He et al., 1996). Subsequently, four additional WAKs were described (AtWAK2-5) sharing similar protein structure and domain compositions, including an N-terminal signal peptide (He et al., 1999). WAK proteins have been shown to be involved in cellular expansion and capable of binding to pectin within the plant cell wall. A homogalacturonan-binding region (later described as a GUB_WAK_bind domain, GWB) was identified near the N-terminal of AtWAK1, facilitating pectin binding (Decreux et al., 2006; Wagner and Kohorn, 2001). Using AtWAK1 as the query in a BLAST search, 22 proteins in *A. thaliana* were found to have high similarity to AtWAKs but lacked some functional domains and were consequently designated as AtWAKLs (Verica and He, 2002). Amongst these, five AtWAKLs were truncated proteins while the rest were transmembrane proteins containing a cytoplasmic kinase domain, an N-terminal signal peptide and extracellular EGF-like or calcium-binding EGF-like domains (some of which were degenerate). The truncated AtWAKLs lacked a transmembrane domain or the EGF-like domains; however, the expression of at least three of these *AtWAKLs* suggests that they may still be functionally important. The definition of a WAK and WAKL based on protein domain composition varies across recent studies, leading to discrepancies in classification.

2.2 Current approaches and inconsistency in the identification and classification of WAK/WAKLs

The main approach for identifying putative WAKs and WAKLs involves using the *A. thaliana* WAK/WAKL protein sequences as queries in a BLASTp (BLAST-protein) search against the proteome of the species of interest. Another method involves constructing a Hidden Markov Model (HMM) profile that includes sequences of the GWB domain-, EGF domain- and STK (or pkinase) domain, usually obtained from the Pfam database. Only candidate proteins (with their corresponding genes) containing both the kinase (STK) domain and the GWB domain are retained for further classification. The verification of protein identity and further classification of these proteins as either a WAK or WAKL are performed using protein domain prediction tools such as SMART, NCBI conserved domain search, and/or IntroPro (Pfam databases). However, inconsistencies in the recent identifications of WAK/WAKLs arise from varying queries, approaches and thresholds. For example, in *Juglans regia* (walnut), only AtWAKs were used in a local BLASTp (LBp) search, whereas in *Rosa chinensis* (rose) both AtWAKs and AtWAKLs were used in an LBp followed by an HMM search with three WAK-expected protein domains (Table 1). Furthermore, differing e-values during the identification of WAK/WAKLs impact the statistical confidence with which a gene is identified as a member of the WAK/WAKL family. For example, potential genes identified in rose with associated e-value between 1e-3 and 1e-5 might be retained in one study but excluded in another using stricter threshold values, such as those used in *Medicago truncatula* (barrel medic) *WAK/WAKL* identification (Kong et al., 2023; Liu et al., 2021).

The criteria for protein domain architecture verification and classification of WAK/WAKLs also vary across different studies. In *Cannabis sativa* (cannabis), all five of the previously described protein domains associated with WAK proteins were used for classification, whereas other studies used different criteria (Sipahi et al., 2022). A study on *Malus domestica* (apple) only investigated the presence of an EGF, TM and pkinase domain to define their WAK proteins (Zuo et al., 2019). Differences in classification criteria typically involve the exclusion of the TM and signal

TABLE 1 Methods and protein domains used for identification and classification of predicted WAK and WAKL proteins across various plant species.

Species (common name)	Identification approach (Query/ e-value)	WAK vs. WAKL: Protein domains used for classification	Reference
Brassica rapa (Chinese cabbage)	LBp & LBn (AtWAK & AtWAKL/< 0.01)	WAK: WAK or GWB &/or WAKa & EGF-like/ EGF_CA & pk & TM	Zhang et al. (2020)
		WAKL: pk &/or WAK/GWB or EGF	
Cannabis sativa (Cannabis)	LBp (AtWAK & AtWAKL/< 1e-5)	WAK: GWB & EGF/EGF_CA & pk & SP & TM	Sipahi et al. (2022)
		WAKL: pk &/or GWB &/or EGF/ EGF_CA &/or SP &/or TM	
Hordeum vulgare (Barley)	HMM + TBLASTn (EGF_CA & pk & GWB/< -100)	N/A	Tripathi et al. (2021)
Juglans mandshurica (Walnut)	LBp (AtWAK/< 1e-5)	WAK: WAK/GWB & STK/PKc, EGF/EGF_CA & SP & TM	Li et al. (2022)
Juglans regia (Walnut)		WAKL: WAK/ GWB & STK/PKc, EGF/EGF_CA &/or SP &/or TM	
Malus domestica (Apple)	Not specified (EGF & pk)	WAK: EGF & TM & pk	Zuo et al. (2019)
Medicago truncatula (Barrel medic)	BLASTp + HMM (AtWAK & OsWAK/< 1e-5)	WAK: EGF_CA & WAKa & GWB & pk_Tyr	Kong et al. (2023)
Nicotiana benthamiana	HMM (AtWAK & AtWAKL & SlWAK & SlWAKL)	WAK: GWB & EGF & pk	Zhong et al. (2023)
		WAKL: GWB &/or EGF &/or pk (Two of three)	
Pisum sativum (Pea)	LBp (AtWAK/< 1e-5)	WAK: STK/ PKc_like & GWB	Li et al. (2023)
Rosa chinensis (Rose)	LBp (AtWAK & AtWAKL) + HMM (EGF_CA, GWB, pk_Ser-Thr/<1e-3)	WAK: EGF_CA, GWB, pk, SP, TM	Liu et al. (2021)
		WAKL: SP, TM, pk &/or EGF_CA and/ or GWB	
Saccharum spontaneum (Wild sugarcane)	HMM (GWB & EGF & pk)	WAK: GWB & EGF & pk	Wang et al. (2023)
Sorghum bicolor (Sorghum)			
Sesame indicum (Sesame)	LBp (AtWAKLs/ $<10^{-10}$)	WAKL: GWB & pk	Yan et al. (2023)
Solanum tuberosum (Potato)	HMM	WAK: GUB_WAK/ WAKa/EGF & TM & pk	Yu et al. (2022)
		WAKL: TM & pk &/or GUB_WAK/ WAKa/EGF	
Triticum aestivum (Bread wheat)	LBp (AtWAK) + HMM (WAK family)	WAK: GWB & TM & pk	Xia et al. (2022)
Zea mays (Maize)	HMM (GWB & EGF_CA & pk/ < 0.001)	WAKL: GWB & EGF_CA &/or pk	Hu et al. (2023)

Gossypium hirsutum (Cotton) and *Solanum lycopersicum* (Tomato) summarized in Figure 1.
HMM, Hidden Markov model search; LBp, Local BLASTp; LBn, Local BLASTn; pk, pkinase; GWB, GUB_WAK_bind; TM, Transmembrane; WAKa, WAK_assoc; STK, STKc_IRAK; PKc, PKc_Superfamily; SP, Signal peptide.
& = must include the previous and following domain.
Or = Must contain either the previous or following domain.

peptide (SP) domains in WAKs. Among studies that classified WAKLs, the criteria also varied from a minimum of a pkinase or GWB domain in *Brassica rapa* (Chinese cabbage) to a minimum of a TM, SP and pkinase domain in rose (Liu et al., 2021; Zhang et al., 2020). Therefore, the WAKs identified in apple and Cannabis, as well as WAKLs from Chinese cabbage and rose are not comparable due to differing criteria, a predicament that applies to many recent characterizations (Table 1). Examples of this inconsistency is also evident in the WAK/WAKL characterization of *Solanum lycopersicum* (tomato) and *Gossypium hirsutum* (cotton) (Figure 1).

Inconsistencies in the methods used for WAK/WAKL identification has led to differences in the *WAK/WAKL* genes identified. For example, two independent characterizations of tomato WAKs yielded different results, with one identifying 16 and the other five WAKs, with only five being shared (Figure 1A; Supplementary Table 1). One difference was the use of different genomes, the *S. lycopersicum* L. SL4.0 genome from an open-source database (https://solgenomics.net/organism/Solanum_lycopersicum/genome/) and the

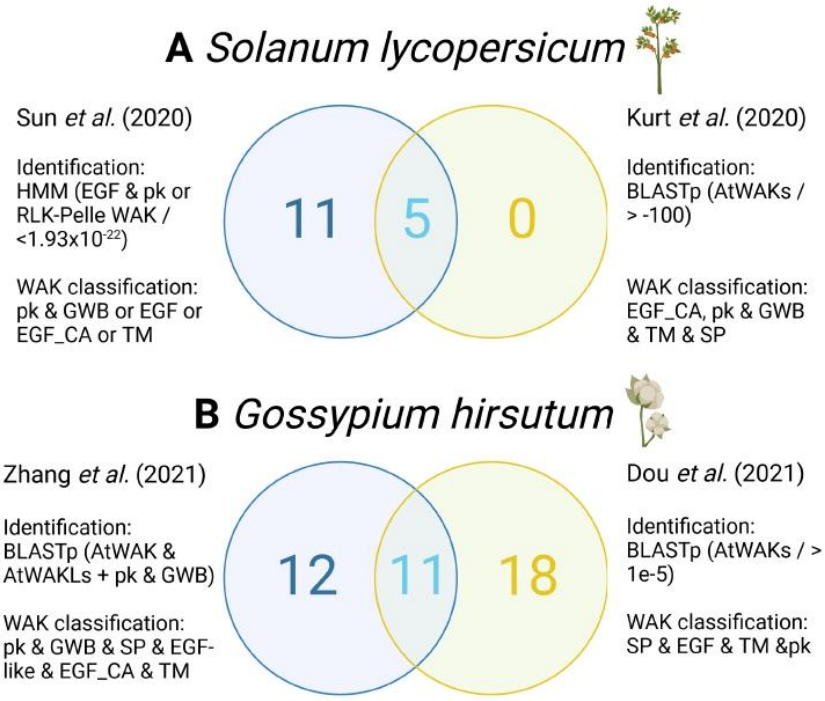


FIGURE 1
Comparison of WAK protein identification and classification methodology and the number and composition of WAK proteins in **(A)** *Solanum lycopersicum* (tomato) and **(B)** *Gossypium hirsutum* (cotton). (Created with BioRender.com).

genome from Phytozome V12.1 (https://phytozome.jgi.doe.gov/pz/portal.html - specific genome and annotation unspecified). The different genomes may have contributed minorly, but a more likely explanation is the use of different identification methods (with different e-value cutoffs) and classification protocols.

The number of cotton WAKs varied considerably between studies, with 11 shared and 12 and 18 unique to each of the respective studies (Figure 1B). Both studies sourced the cotton genome data from the same database (CottonGen - https://www.cottongen.org) with Dou et al. (2021) using the TM-1 genome ZJU_v2.1. and Zhang et al. (2021) likely (genome not specified) using the TM-1 genome UTX_v2.1. Despite both studies using BLASTp for WAK identification they employed different inputs and domains for classification, with one study including a GWB and an additional EGF_CA domain. Even when the identification processes are similar, differences in criteria lead to varying repertoires of WAKs being identified.

To reduce inconsistencies and enable better comparison of this gene family across plant species, future WAK/WAKL characterizations should follow consistent strategies. While identification methods can vary (using BLASTp or HMM searches) the verification and classification process should have the same backbone (such as the inclusion of at least the AtWAK/WAKLs protein sequences as the input query) to minimize the possibility of missed gene members. The verification and classification should however maintain high levels of consistency. Based on the original discovery and descriptions of AtWAK and AtWAKLs, the following model proposes a way to increase standardization: WAK proteins must have an SP, GWB, EGF-like, TM and STK domain in tandem, whereas WAKLs must contain at least a GWB and STK domains with possible varying combinations of the SP, EGF-like and TM domains (Figure 2).

3 *WAK* and *WAKL* composition of species recently characterised

It is important to distinguish between WAK and WAKLs as they could have similar functions, but in different locations in the cell, influencing protein function and response to stimuli. WAKs are defined as transmembrane pectin receptors whereas WAKLs (which are not all transmembrane proteins) should not be classified under the same group. Misclassification could lead to unreliable conclusions as not all members will conform to the same definition. Between 5 and 320 *WAKs* have been identified across various plant species (Table 2). The number of WAKs positively correlates with genome size, but not with ploidy or plant type (monocotyledon or dicotyledon). Bread wheat, a hexaploid, has 320 *WAKs* which is significantly higher than other species (Table 2) (Xia et al., 2022). This could be due to its hexaploid nature, however, this trend is not consistent across other species. For example, the tetraploid potato has fewer *WAKs* (16) compared to cannabis (23), a diploid (Sipahi

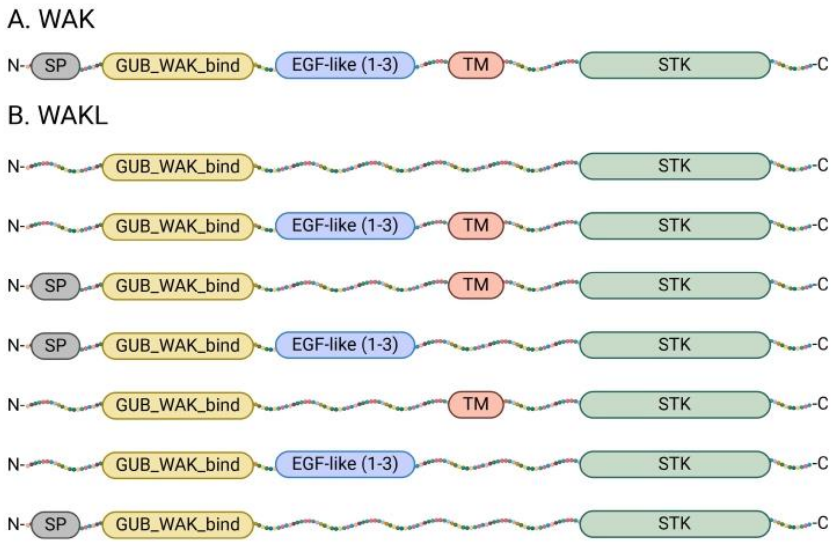


FIGURE 2
Model of required protein domains for the classification of WAK and WAKL proteins. **(A)** Model of the protein domain architecture of WAK proteins. The five domains required are a signal peptide (SP) at the N-terminal (grey), a galacturonan-binding domain (GUB_WAK_Bind; yellow), one to three extracellular epidermal growth factor (EGF)-like domains or EGF_Ca (blue), a transmembrane (TM) domain (Red) and a cytoplasmic Ser/Thr kinase (STK) domain (also known as a pkinase domain at the C-terminal; green). **(B)** Model of the different combinations of protein domains that make up WAKL proteins. They must contain an STK domain and a GUB_WAK_bind domain, but the other domains are variable. (Created with BioRender.com).

et al., 2022; Yu et al., 2022). More research is needed to confirm the correlation between ploidy and the number of *WAK/WAKL* genes. Generally, higher ploidy species tend to have more copies of protein-coding genes but the larger number of *WAKs* could be due to the criteria used for classification. In bread wheat some TaWAKs, such as TraesCS3A02G033400, (and others in phylogenetic group 3), lack an EGF domain and others, like TraesCS7D02G085600, lack an SP which would classify them as WAKLs under the proposed model, thereby reducing the total number of TaWAKs. The less stringent criteria used in the bread wheat study may have contributed to the higher number of TaWAKs identified (Xia et al., 2022).

The number of *WAKL* genes in plant species ranges from 9-82, generally being in higher abundance than the *WAKs* in most species with potato being an exception (Table 2). This could be due to fewer domains required for WAKL protein classification, as fewer domains in coexistence increase the probability of occurrence. This could also result from duplication events of *WAKs*, leading to neofunctionalism, where some domain-coding regions mutate after duplication, allowing for new adaptive functions (Roulin et al., 2013). The exception in potato (16 *StWAKs* vs 13 *StWAKLs*) might be due to different criteria used for StWAKL classification (which lack a TM-domain) unlike other studies. This suggests that potato WAKLs are likely membrane-bound proteins, but other characterised WAKLs lacking a TM domain could be functionally important. Therefore, important potato WAKLs, without TM domains, involved in functions such as pathogen defense might be unidentified and uncharacterised. Not all studies have characterised *WAKLs* in the species studied, thus the data set for *WAKLs* is incomplete. However, *WAKLs* have been shown to play important roles in plants, with AtWAKL10 and OsWAKL21.2 involved in resistance to *Pseudomonas syringae* and *Xanthomonas oryzae*, respectively. This illustrates the value of identifying WAKLs for potential further analyses (Bot et al., 2019; Malukani et al., 2020).

4 Gene comparisons

4.1 *WAK/WAKL* gene structure, placement and duplication methods

The *WAK/WAKL* gene family shows variation in gene structure, with the number of exons ranging from 1-27 exons per gene (Table 3). Some conservation of exon-intron structure (gene structure) exists between monocots and dicots, with most immunity-related WAKs (across plant species) typically having three or four exons, with the first exon being the largest (Stephens et al., 2022). The average number of exons in this family ranges from three to five with open reading frames (ORF) that span between 0.8-330 kb. However, the lack of correlation between ORF length and exon number with gene function (excluding immunity-related WAKs) discounts these properties as a predictive marker for a gene's role. Identifying tandemly duplicated pairs and gene placement is essential for evolutionary analyses to understand the history and potential future of this gene family.

A noticeable feature of the chromosomal placement of these genes is their clustering across multiple chromosomes in species such as Chinese cabbage, rose and *Sesamum indicum* (sesame) (Table 3). This can be indicative of tandem duplication events

TABLE 2 Summary of the number of *WAKs* and *WAKLs* identified in various plant species.

Species (Common name)	Ploidy	Genome size (Mb)	Monocot or Dicot	Number of WAKs	Number of WAKLs	Reference
Brassica rapa (Chinese cabbage)	Allotetraploid	352.8	Dicot	11	85	Zhang et al. (2020)
Cannabis sativa (Cannabis)	Diploid	875.7	Dicot	23	30	Sipahi et al. (2022)
Gossypium arboreum (Cotton)	Diploid	1 621	Dicot	16	42	Zhang et al. (2021)
Gossypium raimondii (Cotton)	Diploid	750.2	Dicot	11	55	
Gossypium hirsutum (Cotton)	Allotetraploid	~2 250-2 430	Dicot	23	76	
Gossypium hirsutum (Cotton)	Allotetraploid	~2 250-2 430	Dicot	29	N/A	Dou et al. (2021)
Hordeum vulgare (Barley)	Diploid	4 226	Monocot	91	N/A	Tripathi et al. (2021)
Juglans mandshurica (Walnut)	Diploid	528.2	Dicot	5	9	Li et al. (2022)
Juglans regia (Walnut)	Diploid	572.8	Dicot	11	16	
Malus domestica (Apple)	Di- or triploid	703	Dicot	44	N/A	Zuo et al. (2019)
Medicago truncatula (Barrel medic)	Diploid	429.6	Dicot	54	N/A	Kong et al. (2023)
Nicotiana benthamiana (Tobacco)	Amphidiploid	N/A	Dicot	15	23	Zhong et al. (2023)
Pisum sativum (Pea)	Diploid	3 796	Dicot	24	N/A	Li et al. (2023)
Rosa chinensis (Rose)	Diploid	515.1	Dicot	23	45	Liu et al. (2021)
Saccharum spontaneum (Wild sugarcane)	Autopolyploid	2 761	Monocot	19	N/A	Wang et al. (2023)
Sorghum bicolor (Sorghum)	Diploid	708.8	Monocot	37	N/A	
Sesame indicum (Sesame)	Diploid	357	Dicot	N/A	31	Yan et al. (2023)
Solanum lycopersicum (Tomato)	Diploid	950	Dicot	11	18	Sun et al. (2020)
Solanum lycopersicum (Tomato)	Diploid	950	Dicot	5	N/A	Kurt et al. (2020)
Solanum tuberosum (Potato)	Tetraploid	705.8	Dicot	16	13	Yu et al. (2022)
Triticum aestivum (bread wheat)	Hexaploid	14 567	Monocot	320	N/A	Xia et al. (2022)
Zea mays (Maize)	Diploid	2 400	Monocot	N/A	58	Hu et al. (2023)

N/A means that that particular data was not discussed in the relevant publication.

occurring, which in turn allows for the expansion of this gene family within a species (Kong et al., 2007). Duplication events involve the replication and insertion of DNA segments into a location close or far away from the original DNA section, producing two copies of a DNA segment. The main duplication mechanisms predicted in the characterization of the *WAK/WAKL* gene families through *in silico* analyses include whole-genome duplications, tandem duplication, proximal duplication and segmental duplication (Lallemand et al., 2020). Tandem duplications were the major type of duplication in Chinese cabbage, walnut, and potato, while, whole-genome-, segmental- and proximal duplications also contributed to *WAK* gene expansion in cotton, Chinese cabbage, *Saccharum spontaneum*, a *Saccharum* hybrid and *Sorghum bicolor* (wild sugarcane and sorghum) (Dou et al., 2021; Wang et al., 2023; Zhang et al., 2020). The type and frequency of duplication events likely depend on the genome organization (ploidy), evolutionary history (recent or distant speciation or hybridization), and current environmental influences (adaptation processes).

TABLE 3 *WAK* and *WAKL* gene information from various species.

Species	Chromosome (chr) placement	Number of exons	Mode number of exons	ORF (Kb)	Clusters [a] (on chr_)	Tandem duplications predicted (%)	K_a/K_s values	Reference
Brassica rapa	On 10/10 chrs	1-27	>3	0.8-20	1, 2, 5, 6, 7, 8	55.17	N/A	Zhang et al. (2020)
Cannabis sativa	On 6/10 chrs	2-7	3	2.5-15	1,4,7	8.7	All < 1 One > 1	Sipahi et al. (2022)
Gossypium arboreum	On 11/13 chrs	N/A	N/A	N/A	N/A	N/A	N/A	Zhang et al. (2021)
Gossypium raimondii	On 11/13 chrs	N/A	N/A	N/A	N/A	N/A	N/A	
Gossypium hirsutum	On 19/23 chrs	1-8	3	2-18	A2, A3, A5, A9, A10, A11, D2, D5, D9, D10, D11	15.15	All < 1	
Gossypium hirsutum	On 12/26 chrs	1-8	3	1-12	A2, A5, D2, D10	17.24	All < 1	Dou et al. (2021)
Juglans mandshurica	On 7/16 chrs	2-8	5	4-330	15	64.29	Most < 1 Two pairs > 1	Li et al. (2022)
Juglans regia	On 8/16 chrs		4		6,10	70.37		
Malus domestica	On 16/17 chrs	1-11	3/4	1.5-5	N/A	N/A	N/A	Zuo et al. (2019)
Medicago truncatula	On 8/8 chrs	2-4	3	2.8-6.5	1,3	18.52	N/A	Kong et al. (2023)
Nicotiana benthamiana	N/A	2-8	>2	1.3-13	N/A	N/A	N/A	Zhong et al. (2023)
Pisum sativum	On 7/7 chrs	1-	>3	2-27	1,3,6	10.71	N/A	Li et al. (2023)
Rosa chinensis	On 7/7 chrs	1-12	3/4	1.8-15.5	1, 2, 5, 7	N/A	N/A	Liu et al. (2021)
Saccharum spontaneum	On 21/32 chrs	1-8	3/4	2.5-25	None	11	All < 1	Wang et al. (2023)
Sorghum bicolor	On 9/10 chrs				4, 7	28		
Sesame indicum	On 8/13 chrs	2-5	3	2-23	3, 5, 6, 8, 10, 11, 12	51.6	N/A	Yan et al. (2023)
Solanum lycopersicum	On 9/12 chrs	1-6	2-3	1-6.5	5, 9, 10	N/A	N/A	Sun et al. (2020)
Solanum lycopersicum	On 2/12 chrs	3-4	3	N/A	7, 9	N/A	N/A	Kurt et al. (2020)
Solanum tuberosum	On 8/24 chrs	1-10	3/4	1.9-10.2	5, 9, 10	72.41	All < 1	Yu et al. (2022)
Zea mays	On 10/10 chrs	2-5	3	3-19.5	1,2,3,6,8	16	All < 1	Hu et al. (2023)

[a]A cluster refers to two or more genes within a few thousand base pairs of each other.
N/A means that that particular data was not discussed in the relevant publication.
ORF, number of exons, mode of exons and clustering estimated off gene structure visualizations when not stated in-text.

For *WAK/WAKL* gene pairs originating from tandem duplication events, nonsynonymous-to-synonymous (K_a/K_s) ratios were calculated to predict the type of evolutionary pressure acting on these pairs. Nonsynonymous mutations (K_a) represent mutations resulting in amino acid changes, while synonymous mutations (K_s) represent mutations resulting in the same amino acid. The K_a/K_s test is an empirical method to predict the type of selection influencing gene evolution, although this method is best suited towards long, single exon genes with relatively low false-positive and false-negative rates compared to other methods (Nekrutenko et al., 2002). Most gene pairs in Table 3 have K_a/K_s ratios below one (<1), indicative of purifying selection, where

deleterious polymorphisms are removed, reducing genetic diversity (Cvijović et al., 2018). This purifying selection likely removes truncated, non-functional proteins while preserving their function. However, one *WAK/WAKL* pair in cannabis and two in walnut show ratios greater than one (>1), indicating positive selection, where advantageous mutations are selected for and become fixed over time (Li et al., 2022; Sipahi et al., 2022; Thiltgen et al., 2017). The hypothesis for the walnut WAK/WAKL pairs under positive selection suggests that variation in the specific duplicated genes, increases the rate of neofunctionalism, allowing for novel or contributory functions to arise (Li et al., 2022). The WAK/WAKLs encoded by these genes may enhance resistance to a specific- or several- or new pathogens or provide developmental advantages.

4.2 *Cis*-acting element comparison in *WAK/WAKL* promotor regions

Cis-acting elements in the promoter regions of genes significantly contribute to differential regulation (Biłas et al., 2016). These elements involved in phytohormone-responsive, light-responsive, biotic- and abiotic-stress responsive, and development-related pathways, enhance regulatory plasticity, allowing genes to be selectively induced or suppressed under certain conditions to perform specific roles (Figure 3).

Elements responsive to developmental pathways are present in the promoter regions of *WAK/WAKLs*, playing roles in leaf cell expansion during developmental stages and potentially in cell tension (Wagner and Kohorn, 2001). Predicted development-related *cis*-acting elements include the AAGAA-motif (endosperm-specific negative expression), AC (xylem-specific expression), O2-site (metabolism of the storage protein, zein) and CAT-box (meristem expression). For example, *CsWAK3* in cannabis contains a root-specific expression element, while *CsWAKL11* contains a seed-specific element, suggesting development-related tissue-specific regulation of these two genes (Sipahi et al., 2022).

Phytohormone-responsive *cis*-acting elements are also present in the promoter regions of many *WAK/WAKLs*, with some genes showing induction following phytohormone treatment. Phytohormones influence plant physiology including growth, development, and abiotic- and biotic stress responses (El Sabagh et al., 2022). Treatment with gibberellin and abscisic acid in potato and cotton induced five *WAKs* and 13 *WAKLs*, potentially due to the presence of ABRE (abscisic acid responsive element), P-box and TATC-box (gibberellin responsive) *cis*-acting elements (Dou et al., 2021; Yu et al., 2022). Other predicted phytohormone-responsive elements in *WAK/WAKL* promoters include those responsive to salicylic acid (TCA-element), methyl jasmonate (TGACG motif), auxin (TGA-box) and ethylene (ERE) in various species such as cotton, apple, potato, tomato and *Nicotiana benthamiana* (Sun et al., 2020; Zhang et al., 2021; Zhong et al., 2023; Zuo et al., 2019).

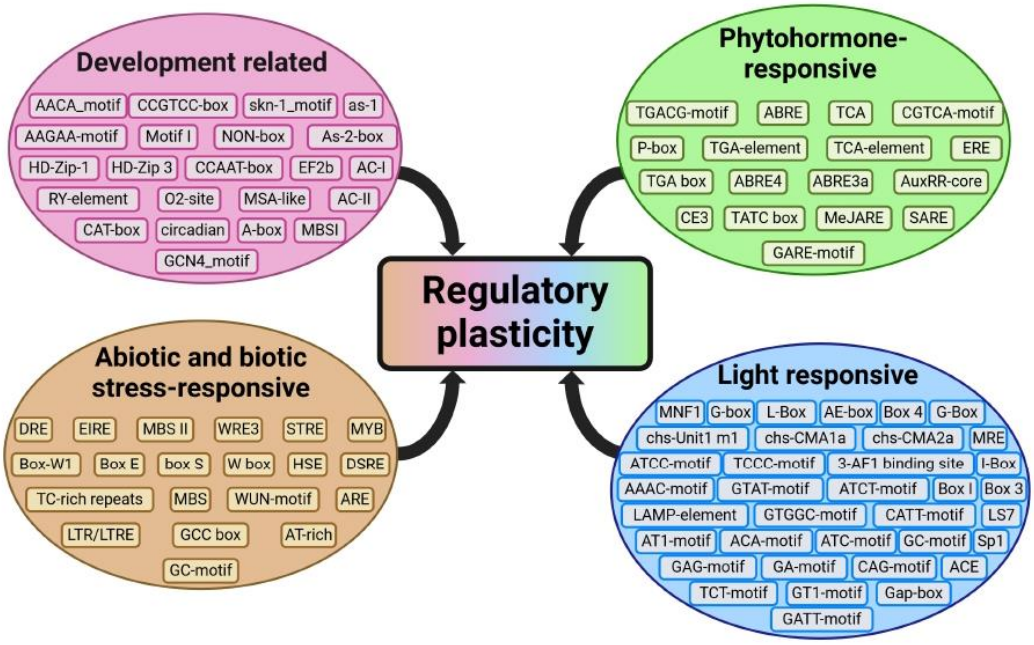


FIGURE 3
Cis-acting elements predicted in the promoter regions (which varied between 1000 and 2000 base pairs upstream of the start site) of *WAK/WAKL* gene families, which contribute to regulatory plasticity. The elements were predicted in walnut, cotton, apple, cannabis, potato, tomato, barley, barrel medic, *Nicotiana benthamiana*, wild sugarcane, and sorghum (Dou et al., 2021; Kong et al., 2023; Li et al., 2022; Sipahi et al., 2022; Tripathi et al., 2021; Wang et al., 2023; Yu et al., 2022; Zhang et al., 2021; Zhong et al., 2023; Zuo et al., 2019). (Created with BioRender.com).

Phytohormones play important roles across many cellular processes and understanding their involvement in these pathways can enhance our knowledge of their function.

Abiotic and biotic stress-responsive elements are implicated in mechanical wounding and defense response pathways in species such as bread wheat, walnut, rose, apple, potato, barrel medic, Chinese cabbage, *N. benthamiana*, and wild sugarcane (Kong et al., 2023; Li et al., 2022; Wang et al., 2023; Xia et al., 2022; Yu et al., 2022; Zhong et al., 2023; Zuo et al., 2019). Examples include the abiotic responsive element, LTR (low-temperature responsive), which regulates gene expression in colder climates and the biotic responsive element, W-box responsive to wounding and pathogen response. Approximately half of the identified tomato *WAK/WAKLs* (14 out of 29) contained the WUN-motif, with additional phytohormone responsive elements (methyl jasmonate, abscisic acid, gibberellin, salicylic acid and, auxin) in the promoter regions. The authors hypothesized that post-wounding rapid induction of *SlWAK-RLKs* may be mediated by these phytohormone pathways, although functional studies are needed (Sun et al., 2020).

In cotton, 13 and 10 *WAK/WAKLs* were induced by gibberellin and auxin, respectively (Dou et al., 2021). Putative gibberellin- and auxin-responsive elements complemented RT-qPCR expression data, linking these *cis*-acting elements directly to gene expression. In potato, gibberellin and auxin induced two and three *WAK/WAKLs*, respectively, with predicted *cis*-acting elements aligning with expression results (Yu et al., 2022). This provides evidence of *WAK*/WAKL involvement in phytohormone pathways and illustrates the relevance of *cis*-acting element predictions. These predictions can guide *in silico* studies regarding *WAK/WAKL* expression patterns in response to phytohormones before conducting expression work, potentially allowing for the editing or selection of promoter regions to control gene expression for specific developmental or defense pathways to maximize efficiency.

Light-responsive elements were predicted in apple, tomato and walnut, with common elements such as Box 4 and the TCT-motif (Li et al., 2022; Sun et al., 2020; Zuo et al., 2019). These elements may facilitate a 'circadian clock' of *WAK/WAKL* genes, as seen in other receptor-like kinases (Zuo et al., 2019). This suggests a time-specific induction or suppression of certain *WAK/WAKLs* to support time-sensitive cellular functions, potentially impacting growth patterns and allowing cellular expansion towards the light during the day.

While the results from these studies provide a foundation for comparison, discrepancies in prediction methods may impact general conclusions. Typically, *cis*-acting elements are predicted in the region 2000 bp upstream of a gene's transcriptional start site, but this was not followed for apple and tomato, where 1500 bp promoters were used (Sun et al., 2020; Zuo et al., 2019). Reduced promoter size impacts the abundance of predicted elements, making comparisons difficult. Consistency in promoter length within the same gene family is crucial for comparability. The lack of e-value cutoffs for element prediction further complicates the comparison. Standardizing prediction methods would improve the comparability of this gene family across species, enabling larger-scale conclusions and more accurate analyses. Identifying common elements across species could provide research targets for studying gene regulation and function, potentially leading to overexpression or suppression of genes of interest in future studies or genetically modified organisms (GMOs) for desired outcomes.

5 Expression data comparison

Inferences about functional roles are made by evaluating gene expression patterns (here focusing on *WAK/WAKLs*). These studies can indicate the involvement of genes in specific tissues, phytohormone responses, abiotic stresses (such as drought and cold), and biotic stresses (such as pathogen infection). Upregulation during environmental changes is a primary line of evidence for implicating genes in particular responses. For example, cotton *WAK/WAKLs* were upregulated in response to cold, heat, salt, and chemically-induced drought stress suggesting their involvement, while upregulated barrel medic *WAK/WAKLs* were involved in the defense responses against *Macrophomina phaseolina* and *Ralstonia solanacearum* (Kong et al., 2023; Zhang et al., 2021). Expression data are typically obtained from RT-qPCR experiments or RNA-sequencing data, which is often validated with RT-qPCR. Unlike previous sections, expression analysis approaches were more consistent across various studies, increasing the reliability of the conclusions drawn.

5.1 Tissue specificity and developmental stages

Phylogenetic group-specific tissue expression was observed when assessing the Chinese cabbage *BrWAK/WAKLs*, with three genes expressed in roots, five in flowers and three in calluses (Zhang et al., 2020). This suggests a close correlation between *BrWAK/WAKL* evolutionary relationships and tissue specificity. Tissue-specific expression was assessed in cannabis, sesame, *N. benthamiana*, walnut and bread wheat in tissues such as roots, stems, leaves, pericarp, grains, spikes, buds, pods, seeds and flowers (Kong et al., 2023; Li et al., 2022; Sipahi et al., 2022; Xia et al., 2022; Yan et al., 2023; Zhong et al., 2023). In barrel medic, some *MtWAKs*, like *MtWAK24* and *MtWAK50* showed expression specifically in roots, while others, such as *MtWAK7* and *MtWAK8* were expressed in multiple tissues (Kong et al., 2023). These results indicate that *WAK/WAKL* genes can be expressed in a few or many different plant tissues simultaneously or exhibit tissue-specific expression.

Expression data during different developmental stages was obtained for cotton, showing the involvement of a subset of *WAK/WAKLs* in the initial developmental stages of fiber development, and another subset in the cellular elongation phase (Dou et al., 2021). In tomato, expression data for *WAK/WAKLs* shows the involvement of a subset of *SlWAK/SlWAKLs* in fruit development, with varying expression patterns during fruit expansion and ripening (Sun et al., 2020). Identifying and validating genes involved in functions such as ripening could aid

in the selection of varieties during screening and breeding programs to produce larger fruits with controlled ripening times. Most of the barrel medic's *MtWAKs* were downregulated during nodulation, except for the induction of *MtWAK1*, suggesting its contribution to the process (Kong et al., 2023). The *WAK/WAKL* repertoire can thus play a role in many different developmental stages and illustrates the specificity of function some *WAK/WAKLs* have.

5.2 Abiotic and biotic responses

When characterizing the *WAK/WAKL* gene family, one major focus is on linking gene subsets with abiotic and biotic stress responses. In cotton, expression data showed that *GhWAK/WAKLs* responded to abiotic factors including cold, heat, salt, and polyethylene glycol (drought simulation). Specifically, *GhWAK9* was induced in all four abiotic responses, while *GhWAKL17* was only involved in the cold response (Zhang et al., 2021). This demonstrates that *WAK/WAKLs* can have a universal or specific role during stress. Similarly, in Chinese cabbage, expression patterns during drought, high temperature, and high humidity stress showed an association of *BrWAK/WAKLs* in these abiotic responses (Zhang et al., 2020). In tomato, nine *WAK/WAKLs* were upregulated and three downregulated in response to mechanical wounding, indicating their role in the wounding response (Sun et al., 2020). In barrel medic, 10 *MtWAKs* were upregulated due to drought, seven due to cold and three due to salt stress, suggesting that a combined 12 *MtWAKs* are involved in multiple abiotic stress responses (Kong et al., 2023). These findings illustrate that WAK/WAKLs play significant roles in various abiotic stress responses, with specific genes involved in specific responses.

A subset of *WAK/WAKLs* have been associated with biotic defense responses through gene expression analyses, with some functionally shown to play a role in defense. *AtWAKL22* and *TaWAK2* are known to protect against *Fusarium oxysporum* and *Fusarium graminearum*, respectively (Diener and Ausubel, 2005; Gadaleta et al., 2019; Liu et al., 2021; Zhong et al., 2023). Upregulation during defense responses implicated 20 *WAK/WAKLs* in bread wheat, eight in walnut, 12 in rose, eight in sesame, and six in potato (Li et al., 2022; Liu et al., 2021; Xia et al., 2022; Yan et al., 2023; Yu et al., 2022). In barrel medic, nine out of 12 *MtWAKs* were induced by yeast elicitor treatment, one by *M. phaseolina* (necrotrophic fungus) infection, three by *R. solanacearum* (bacteria) infection and 10 by *Erysiphe pisi* (biotrophic fungus) (Kong et al., 2023). Different *WAK/WAKLs* were upregulated following *M. phaseolina* and *R. solanacearum* infection, suggesting pathogen-specific roles in the respective defense responses. Overexpression of *MtWAK24* in *N. benthamiana* leaves reduced lesion size following *Phytophthora parasitica* infection, validating its important role in the inhibition of *P. parasitica* infection (Kong et al., 2023).

Some studies initially used *in silico* expression data to suggest a gene's involvement in defense and then confirmed some genes with functional work, thus validating the *in silico* predictions. Virus induced gene silencing (VIGS) was used in rose and *N. benthamiana* to silence *WAK/WAKLs* involved in defense, which increased disease severity. Silencing *RcWAK4* in rose enhanced susceptibility to *Botrytis cinerea* while silencing *NbWAK12* and *14* and *NbWAKL6* and *12* in *N. benthamina* increased susceptibility to tomato yellow curl leaf virus (Liu et al., 2021; Zhong et al., 2023). This molecular evidence substantiates the defense role of the selected *WAK/WAKLs*. While *in silico* predictions are the first step towards identifying potential defense-related genes, molecular validation (such as VIGS or other functional work) is essential to verify these predictions, ensuring accurate identification of *WAK/WAKLs* involved in stress responses before utilizing them in crop improvement.

6 Protein comparisons

6.1 General properties and subcellular localization

The recently characterised WAK/WAKL proteins across species range between 200-3350 amino acids in length, 22.3-155.59 kDa in weight and 4.86-9.4 isoelectric points (Table 4). These proteins are larger with properties in the normal ranges for plant proteins (Mohanta et al., 2019; Ramírez-Sánchez et al., 2016). WAK proteins likely localize to the plasma membrane due to the presence of a transmembrane domain. However, WAKL proteins do not always contain a transmembrane domain, resulting in variable cellular localizations. *In silico* characterization of the WAK/WAKL family has predicted localizations to various other cellular structures, including the nucleus, chloroplast, extracellular matrix, mitochondria, endoplasmic reticulum, golgi body, vacuoles and the cytoplasm. Some of these localizations have been confirmed through experiments utilizing green fluorescent protein (GFP) markers (Figure 4), indicating that the WAK/WAKLs (specifically the WAKLs) can function throughout the cell, not just at the plasma membrane.

Molecular confirmation of *in silico* predictions was achieved in cotton and barrel medic using GFP markers to demonstrate the subcellular localization of specific WAKs (Kong et al., 2023; Zhang et al., 2021). Five bread wheat *TaWAKs* fused with a *GFP* gene were transiently expressed in wheat protoplasts. Four of the five showed fluorescence only at the plasma membrane, while the GFP-TraesCS3D02G046900 showed fluorescence throughout the protoplast, with higher intensity at the plasma membrane and the nucleus (Xia et al., 2022). The authors suggested that the plasma membrane localization allows these proteins, in part, to act as receptors during immune responses, and the localization of one WAK to the nucleus warrants further investigation. These findings imply that TraesCS3D02G046900 functions by detecting DAMPs in the cytoplasm rather than the extracellular space, thereby signaling successful pathogen penetration that influences the expression of a different repertoire of defense-related genes compared to other WAKs. This particular protein sequence (TraesCS3D02G046900.1), identified as a WAK, contains the pkinase, GUB_WAK_bind, SP and TM domains but lacks an EGF domain, classifying it as a WAKL by earlier definitions (Figure 2). The bread wheat online annotation reveals eight different ORFs for this gene with some producing

TABLE 4 WAK and WAKL proteins characterised between the year 2018-2023.

Species	WAK/WAKL protein length (aa)	WAK/WAKL protein weight (kDa)	WAK/WAKL protein pI	Reference
Brassica rapa	~200-3350	N/A	N/A	Zhang et al. (2020)
Cannabis sativa	582-983	65,6-108,8	5,80-8,96	Sipahi et al. (2022)
Gossypium arboreum	302-1049	48,86-117,73	5,00-9,20	Zhang et al. (2021)
Gossypium raimondii				
Gossypium hirsutum				
Gossypium hirsutum	606-1200	67.36-134.02	5.11-8.79	Dou et al. (2021)
Juglans mandshurica	629-1396	69.19-155.95	4,86-8,83	Li et al. (2022)
Juglans regia	513-839	57.57-92.13		
Malus domestica	302-998	33,63-110,53	5,1-9,26	Zuo et al. (2019)
Medicago truncatula	52-771	5.55-85.6	4.47-10.17	Kong et al. (2023)
Nicotiana benthamiana	202-1159	22,3-128,2	4,9-9,4	Zhong et al. (2023)
Pisum sativum	623-754	70.38-83.26	5.14-8.93	Li et al. (2023)
Rosa chinensis	369-891	N/A	N/A	Liu et al. (2021)
Saccharum spontaneum	441-1710	47,39-126,53	N/A	Wang et al. (2023)
Sorghum bicolor				
Sesame indicum	504-787	57.3-86.62	5.37-8.83	Yan et al. (2023)
Solanum lycopersicum	302-663	42.9-74.9	5.76-8.85	Sun et al. (2020)
Solanum lycopersicum	732-799	81.21-88.30	6.13-8.39	Kurt et al. (2020)
Solanum tuberosum	534-1045	59,98-116,77	5,20-8,33	Yu et al. (2022)
Triticum aestivum	N/A	65,4-119,2	5,09-9,24	Xia et al. (2022)
Zea mays	342-1328	38.03-147.35	4.96-8.96	Hu et al. (2023)

N/A means that that particular data was not discussed in the relevant publication.

proteins lacking the TM domain (http://plants.ensembl.org/Triticum_aestivum/Gene/Summary?g=TraesCS3D02G046900) which could explain the dispersal of this protein throughout the cell. This emphasizes the need for a consistent definition of WAKs and WAKLs to improve comparability across this gene family to enhance our understanding of their function within the plant cell. Subcellular localization predictions provide a foundation for further molecular confirmation, because knowing where proteins function can inform their roles and downstream processes and the cell structures they influence.

6.2 Phylogenetic analyses for functional inference and conserved motif analyses

Phylogenetic analyses of newly identified WAK/WAKL protein sequences are constructed to assess the evolutionary relationships and to infer functions based on similarities to other known protein sequences (Sjölander, 2004). These phylogenetic analyses include the species-specific protein sequences or include previously characterised protein sequences from other species. The WAK/WAKLs typically form three to six phylogenetic clusters depending on the study. Despite using protein sequences, these clusters often reflect different gene structures, such as the number and placement of exons within the ORF. This has been observed in *N. benthamiana*, potato, tomato, and cannabis (Sipahi et al., 2022; Sun et al., 2020; Yu et al., 2022; Zhong et al., 2023). Phylogenetic clustering of proteins with similar gene structures provides an additional level of evidence for the role of tandem duplications within species, supporting the idea of gene family expansions within these species. In sorghum, however, the number of exons varies within groups, which suggests a different evolutionary process for the sorghum *WAK/WAKL* genes (Wang et al., 2023). Moreover, genes from different species with similar structures are often more closely related to each other than to those genes with differing structures within the same species. Species-specific clades in the *WAK/WAKL* gene family of walnut and cotton limit the ability to make broader functional inferences (Li et al., 2022; Zhang et al.,

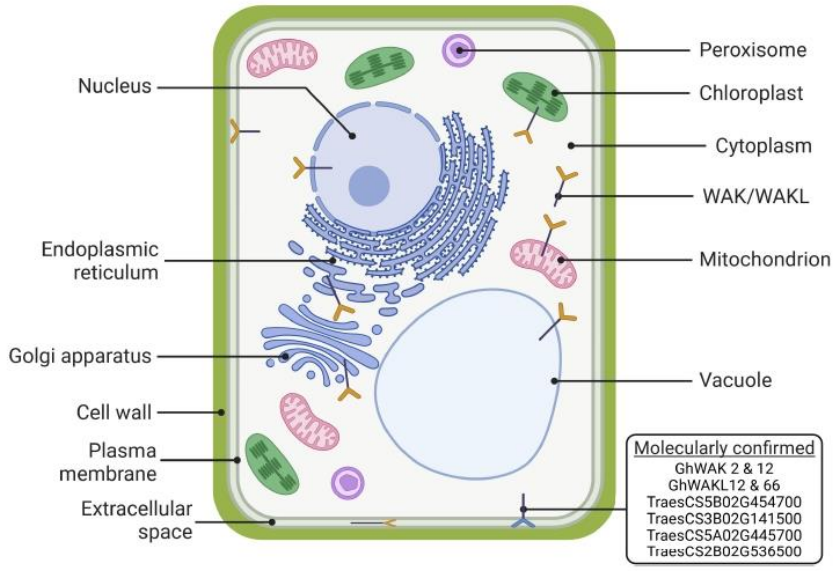


FIGURE 4
Subcellular localization predicted by *in silico* methods (yellow icons) or demonstrated through localization experiments (blue icon) of recently characterised WAK and WAKL proteins in bread wheat, walnut, cotton, apple, cannabis, pea, barrel medic, and potato (Kong et al., 2023; Li et al., 2022, 2023; Sipahi et al., 2022; Xia et al., 2022; Yu et al., 2022; Zhang et al., 2021; Zuo et al., 2019). (Created with BioRender.com).

2021). The *Arabidopsis WAKs* are well studied, but often form independent clusters in phylogenetic analyses, hampering cross-species comparisons. This suggests that the *WAK/WAKL* gene family expands and evolves independently within each species, becoming more distinct over time (Zhang et al., 2021).

The composition of sequences used for phylogenetic analyses varies between studies. Some only include the species' identified WAK/WAKLs, others incorporate *A. thaliana* sequences (including studies done in bread wheat and tomato) and some (potato and cannabis) include functionally characterised WAK/WAKLs involved in development, cell elongation and defense from various species such as *A. thaliana*, rice, cotton, maize, tomato, and apple (Sipahi et al., 2022; Sun et al., 2020; Xia et al., 2022; Yu et al., 2022). These analyses serve different purposes but often provide varying levels of characterization. A recent detailed phylogenetic analysis of 1061 *WAK* genes from 37 species revealed five clades, providing evidence of lineage-specific expansion after speciation, consistent with smaller-scale analyses (Zhang et al., 2023). The main purpose of these analyses is to assess relationships between previously characterised sequences, infer function, or predict the evolution of the gene family within the species. Including as many functionally characterised WAK/WAKLs as possible maximizes the potential for inferring function, especially since *A. thaliana* sequences often cluster independently, as seen in sesame, walnut and cotton WAK/WAKLs (Li et al., 2022; Yan et al., 2023). If the *WAK/WAKLs* of only the species of interest are included, inferences could be made on the evolution and expansion of the gene family within the species, allowing for a hypothesized description of the gene family's history. Consistency in the approach would enhance comparability among the gene family as similar levels of description will be available.

When assessing conserved protein motifs, most characterizations use the online, open-source platform MEME (Multiple EM for Motif Elicitation - https://meme-suite.org/meme/tools/meme) (Bailey et al., 2015). MEME identifies near-exact repeating patterns of sequences (conserved motifs) with statistical modelling. The maximum number of conserved motifs is manually set, representing the best statistical set. To ensure comparability, the number of conserved motifs should be constant, ideally at 10, so that only the 10 most significant motifs can be compared between WAK/WAKLs from different species. Only cannabis, tomato and barley studies provided motif sequences, showing little similarity in motif compositions (Sipahi et al., 2022; Sun et al., 2020; Tripathi et al., 2021).

Ten motifs were identified in cannabis, potato, tomato, barley and *Saccharum* species, 12 in cotton, and 15 in *N. benthamiana* (Sipahi et al., 2022; Sun et al., 2020; Tripathi et al., 2021; Wang et al., 2023; Yu et al., 2022; Zhang et al., 2021; Zhong et al., 2023). Groupings obtained by phylogenetic analysis often align with motifs as seen in the group II cluster of potato WAKLs (Yu et al., 2022). *N. benthamiana* and *Saccharum* spp. WAK/WAKLs contain motifs specific to phylogenetic groups, with higher similarity within groups than those between groups (Wang et al., 2023; Zhong et al., 2023). The distribution and types of motifs in WAK/WAKLs are largely conserved within phylogenetic groups likely due to the absence of exon -gain and -loss over long evolutionary periods, indicating functional conservation (Sun et al., 2020; Yu et al., 2022; Zhong et al., 2023). In cotton, variations were mainly observed in the N-terminal of the proteins, where the SP, GUB_WAK_bind and EGF-Like domains are located (Zhang et al., 2021). Conserved

domains were primarily in the C-terminal (STK domain), suggesting a highly conserved kinase domain for downstream functioning. Less conservation in the SP, the pectin binding and EGF-like domains could allow differential translocation between proteins, recognition of different pectin forms, and interaction with a wide range of additional proteins, respectively. Standardizing the number of predicted motifs at 10 will allow more direct comparisons across species, ensuring the best statistical set of representative motifs.

6.3 Impact of polymorphism on function

Polymorphisms in the WAK sequences can significantly affect protein functioning. The *RFO1* gene (annotated as *AtWAKL22*), is a key defense gene in *A. thaliana*, providing resistance against *Fusarium oxysporum* infection and was later described to detect changes in methylation status of cell wall pectin following infection (Huerta et al., 2023). Polymorphisms between ecotypes of this gene influence the protein's efficiency in the defense response (Diener and Ausubel, 2005). The authors identified 21 polymorphisms between two ecotypes, including two 3 bp deletions as well as 10 causing missense mutations. One mutation changed the highly conserved glutamic acid (nucleotide 1652) in the resistant ecotype to glutamine in the susceptible ecotype. Similar sequence differences were seen in the *WAK2* genes of resistant and susceptible wheat, including a deletion, an insertion, and an A-to-C substitution (Gadaleta et al., 2019). Three independent *wak2* mutant wheat lines, with predicted loss-of-function mutations, exhibited severe disease symptoms. The authors hypothesized that these polymorphisms, along with alternative splicing, form different WAK protein variants, influencing the efficiency of the defense response. Focusing on the effect of polymorphisms within WAK/WAKLs among members of a species adapted to different biotic and abiotic stresses is important. Identifying these polymorphisms could improve *in silico* characterizations as an additional predictive tool for functional inferences. However, creating a comprehensive database of all known polymorphic regions with associated functionalities is essential before this type of *in silico* analysis can be implemented.

7 Discussion

The *WAK/WAKL* gene family has attained significant interest recently, leading to a substantial influx of foundational characterization studies. However, inconsistencies in the methodologies for identification, classification, and characterization of this gene family

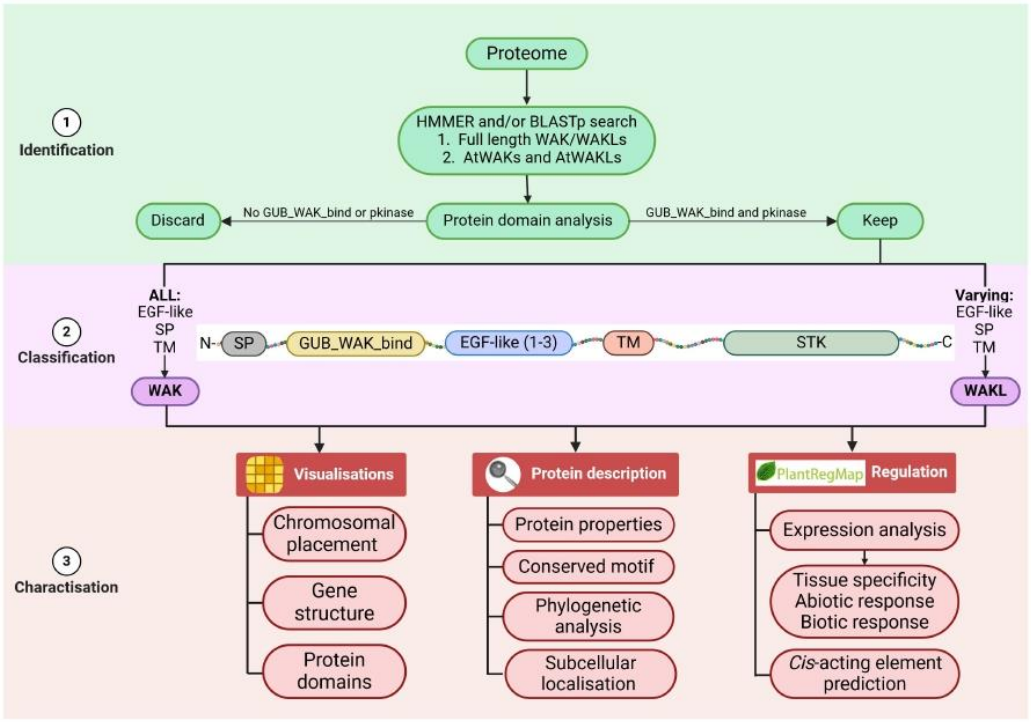


FIGURE 5
Proposed methodology for identifying, classifying and characterizing the *WAK/WAKL* gene family using species-specific proteomes. The outline specifies the identification methods, the minimum inputs required for classification, and which sequences should be excluded. Following classification, it lists the general analyses to be performed for a comprehensive *in silico* characterization of the gene family. (Created with BioRender.com).

have hampered the comparability and robustness of the results. Figure 5 provides a guideline to standardize the identification, classification, and characterization of *WAK/WAKL* gene families, aiming to streamline the process and improve consistency across different species.

The identification process begins with the predicted proteome of the species of interest. This proteome is used in a HMMER or BLASTp search utilizing full-length WAK/WAKL sequences, with at least the AtWAK/WAKL sequence, as the query. Using full-length proteins as inputs is crucial since domain-based HMMER searches may overlook genes lacking specific queried domains, given the previous variations in WAKL definitions. Only proteins identified with an e-value < $1e^{-5}$ should be considered for further investigation, as this is the typical threshold used. Proteins not containing both the GUB_WAK_bind and pkinase domains (or have the domains but with an associated e-value > $1e^{-5}$) should be discarded, while those that contain both domains are kept for classification. The classification of WAKs and WAKLs should follow the model proposed in this review, with all domains having an e-value < $1e^{-5}$ (Figure 2). Once classified, the *WAK/WAKL* genes can undergo thorough characterization to fully describe the family. This characterization provides information on the genes, their protein products, and their regulation for functional implication. Visualizations can be created using TBtools with information extracted from general feature format (gff or gff3) files for gene structure and chromosomal placement, and output from the NCBI CDD search for protein domains (Chen et al., 2020; Marchler-Bauer et al., 2011). Protein descriptions can be done by extracting the protein sequences and using tools such as Genefinity (http://www.geneinfinity.org/) for protein property predictions, MEME (https://meme-suite.org) for conserved motif prediction, Genious Prime or MEGA for the phylogenetic analysis using protein sequences and WoLF PSORT (https://wolfpsort.hgc.jp/) for subcellular localization prediction (Bailey et al., 2006; Horton et al., 2007; Tamura et al., 2021). Regulation can be assessed by visualizing expression data generated (or obtained from an online database) with R-studio or TBtools, while *cis*-acting elements can be predicted with PlantRegMap (https://plantregmap.gao-lab.org/) and/or PlantCARE (https://bioinformatics.psb.ugent.be/webtools/plantcare/html/) with *A. thaliana* as the reference species (Chen et al., 2020; Lescot et al., 2002; Racine, 2012; Tian et al., 2019).

By implementing this model for future characterizations, there will be an increased consistency across the *WAK/WAKL* gene family, facilitating better comparisons between and across species. To address previously described *WAK/WAKL*, the coding sequences should be made available for future researchers to confirm or adjust annotations using these guidelines for more accurate large-scale analyses. Both WAKs and WAKLs have been shown to be involved in various developmental and defense-related pathways, playing important roles in plants. Incorrectly identifying a WAKL as a WAK, or identifying a WAKL when it does not meet the minimum criteria, can impact future functional work looking into their biological significance, as there may be differences in modes of functioning due to differing defining features such as functional protein domains. Improved characterization and comparisons of gene families will enhance bioinformatic predictions based on previous data, leading to a better selection of candidate genes for functional analysis.

Author contributions

AH: Conceptualization, Data curation, Formal analysis, Methodology, Visualization, Writing – original draft, Writing – review & editing. NV: Funding acquisition, Resources, Supervision, Writing – review & editing. VS: Conceptualization, Funding acquisition, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. Funding was provided by the NRF and the Hans Merensky Legacy Foundation.

Acknowledgments

The authors would like to acknowledge the members of the Hans Merensky Chair in Avocado Research for their support.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2024.1467148/full#supplementary-material

References

Bailey, T. L., Johnson, J., Grant, C. E., and Noble, W. S. (2015). The MEME suite. *Nucleic Acids Res.* 43, W39–W49. doi: 10.1093/nar/gkv416

Bailey, T. L., Williams, N., Misleh, C., and Li, W. W. (2006). MEME: Discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* 34, W369–W373. doi: 10.1093/nar/gkl198

Biłas, R., Szafran, K., Hnatuszko-Konka, K., and Kononowicz, A. K. (2016). *Cis*-regulatory elements used to control gene expression in plants. *Plant Cell Tissue Organ Culture (PCTOC)* 127, 269–287. doi: 10.1007/s11240-016-1057-7

Bot, P., Mun, B.-G., Imran, Q. M., Hussain, A., Lee, S.-U., Loake, G., et al. (2019). Differential expression of *AtWAKL10* in response to nitric oxide suggests a putative role in biotic and abiotic stress responses. *PeerJ* 7, e7383. doi: 10.7717/peerj.7383

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: An integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 13, 1194–1202. doi: 10.1016/j.molp.2020.06.009

Cvijović, I., Good, B. H., and Desai, M. M. (2018). The effect of strong purifying selection on genetic diversity. *Genetics* 209, 1235–1278. doi: 10.1534/genetics.118.301058

Decreux, A., Thomas, A., Spies, B., Brasseur, R., Cutsem, P. V., and Messiaen, J. (2006). *In vitro* characterization of the homogalacturonan-binding domain of the wall-associated kinase WAK1 using site-directed mutagenesis. *Phytochemistry* 67, 1068–1079. doi: 10.1016/j.phytochem.2006.03.009

Diener, A. C., and Ausubel, F. M. (2005). *Resistance to fusarium oxysporum 1*, a dominant *Arabidopsis* disease-resistance gene, is not race specific. *Genetics* 171, 305–321. doi: 10.1534/genetics.105.042218

Dou, L., Li, Z., Shen, Q., Shi, H., Li, H., Wang, W., et al. (2021). Genome-wide characterization of the *WAK* gene family and expression analysis under plant hormone treatment in cotton. *BMC Genomics* 22, 1–17. doi: 10.1186/s12864-021-07378-8

El Sabagh, A., Islam, M. S., Hossain, A., Iqbal, M. A., Mubeen, M., Waleed, M., et al. (2022). Phytohormones as growth regulators during abiotic stress tolerance in plants. *Front. Agron.* 4. doi: 10.3389/fagro.2022.765068

Gadaleta, A., Colasuonno, P., Giove, S. L., Blanco, A., and Giancaspro, A. (2019). Map-based cloning of QFhb. mgb-2A identifies a *WAK2* gene responsible for Fusarium Head Blight resistance in wheat. *Sci. Rep.* 9, 1–15. doi: 10.1038/s41598-019-43334-z

He, Z.-H., Cheeseman, I., He, D., and Kohorn, B. D. (1999). A cluster of five *cell wall-associated receptor kinase* genes, *Wak1–5*, are expressed in specific organs of *Arabidopsis*. *Plant Mol. Biol.* 39, 1189–1196. doi: 10.1023/A:1006197318246

He, Z.-H., Fujiki, M., and Kohorn, B. D. (1996). A Cell wall-associated, receptor-like protein kinase. *J. Biol. Chem.* 271, 19789–19793. doi: 10.1074/jbc.271.33.19789

Horton, P., Park, K.-J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. J., et al. (2007). WoLF PSORT: protein localization predictor. *Nucleic Acids Res.* 35, W585–W587. doi: 10.1093/nar/gkm259

Hu, K., Dai, Q., Ajayo, B. S., Wang, H., Hu, Y., Li, Y., et al. (2023). Insights into *ZmWAKL* in maize kernel development: genome-wide investigation and GA-mediated transcription. *BMC Genomics* 24, 760. doi: 10.1186/s12864-023-09849-6

Huerta, A. I., Sancho-Andrés, G., Montesinos, J. C., Silva-Navas, J., Bassard, S., Pau-Roblot, C., et al. (2023). The WAK-like protein RFO1 acts as a sensor of the pectin methylation status in *Arabidopsis* cell walls to modulate root growth and defense. *Mol. Plant* 16, 865–881. doi: 10.1016/j.molp.2023.03.015

Kohorn, B., and Kohorn, S. (2012). The cell wall-associated kinases, WAKs, as pectin receptors. *Front. Plant Sci.* 3. doi: 10.3389/fpls.2012.00088

Kohorn, B. D. (2015). Cell wall-associated kinases and pectin perception. *J. Exp. Bot.* 67, 489–494. doi: 10.1093/jxb/erv467

Kong, H., Landherr, L. L., Frohlich, M. W., Leebens-Mack, J., Ma, H., and Depamphilis, C. W. (2007). Patterns of gene duplication in the plant *SKP1* gene family in angiosperms: evidence for multiple mechanisms of rapid gene birth. *Plant J.* 50, 873–885. doi: 10.1111/j.1365-313X.2007.03097.x

Kong, W., Shi, J., Yang, B., Yu, S., Zhao, P., Guo, Z., et al. (2023). Genome-wide analysis of the *wall-associated kinase* (*WAK*) genes in *Medicago truncatula* and functional characterization of *MtWAK24* in response to pathogen infection. *Plants (Basel)* 12, 1849. doi: 10.3390/plants12091849

Kurt, F., Kurt, B., and Filiz, E. (2020). Wall associated kinases (*WAKs*) gene family in tomato (*Solanum lycopersicum*): Insights into plant immunity. *Gene Rep.* 21, 100828. doi: 10.1016/j.genrep.2020.100828

Lallemand, T., Leduc, M., Landès, C., Rizzon, C., and Lerat, E. (2020). An overview of duplicated gene detection methods: Why the duplication mechanism has to be accounted for in their choice. *Genes (Basel)* 11, 1046. doi: 10.3390/genes11091046

Lescot, M., Déhais, P., Thijs, G., Marchal, K., Moreau, Y., Van De Peer, Y., et al. (2002). PlantCARE, a database of plant *cis*-acting regulatory elements and a portal to tools for in *silico* analysis of promoter sequences. *Nucleic Acids Res.* 30, 325–327. doi: 10.1093/nar/30.1.325

Li, M., Ma, J., Liu, H., Ou, M., Ye, H., and Zhao, P. (2022). Identification and characterization of *wall-associated kinase* (*WAK*) and *WAK-like* (*WAKL*) gene family in *juglans regia* and its wild related species *juglans mandshurica*. *Genes* 13, 134. doi: 10.3390/genes13010134

Li, X., Ou, M., Li, L., Li, Y., Feng, Y., Huang, X., et al. (2023). The *wall-associated kinase* gene family in pea (*Pisum sativum*) and its function in response to B deficiency and Al toxicity. *J. Plant Physiol.* 287, 154045. doi: 10.1016/j.jplph.2023.154045

Liu, X., Wang, Z., Tian, Y., Zhang, S., Li, D., Dong, W., et al. (2021). Characterization of *wall-associated kinase/wall-associated kinase-like* (*WAK/WAKL*) family in rose (*Rosa chinensis*) reveals the role of *RcWAK4* in *Botrytis* resistance. *BMC Plant Biol.* 21, 1–12. doi: 10.1186/s12870-021-03307-9

Malukani, K. K., Ranjan, A., Hota, S. J., Patel, H. K., and Sonti, R. V. (2020). Dual activities of receptor-like kinase OsWAKL21.2 induce immune responses. *Plant Physiol.* 183, 1345–1363. doi: 10.1104/pp.19.01579

Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., Deweese-Scott, C., et al. (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* 39, D225–D229. doi: 10.1093/nar/gkq1189

Mohanta, T. K., Khan, A., Hashem, A., Abd_Allah, E. F., and Al-Harrasi, A. (2019). The molecular mass and isoelectric point of plant proteomes. *BMC Genomics* 20, 631. doi: 10.1186/s12864-019-5983-8

Nekrutenko, A., Makova, K. D., and Li, W. H. (2002). The $K(_a)/K(_s)$ ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res.* 12, 198–202. doi: 10.1101/gr.200901

Racine, J. S. (2012). R studio: a platform-independent IDE for R and sweave. *J. Appl. Econometrics* 27, 167–172. doi: 10.1002/jae.v27.1

Ramirez-Sánchez, O., Pérez-Rodriguez, P., Delaye, L., and Tiessen, A. (2016). Plant proteins are smaller because they are encoded by fewer exons than animal proteins. *Genomics Proteomics Bioinf.* 14, 357–370. doi: 10.1016/j.gpb.2016.06.003

Roulin, A., Auer, P. L., Libault, M., Schlueter, J., Farmer, A., May, G., et al. (2013). The fate of duplicated genes in a polyploid plant genome. *Plant J.* 73, 143–153. doi: 10.1111/tpj.12026

Sipahi, H., Whyte, T. D., Ma, G., and Berkowitz, G. (2022). Genome-wide identification and expression analysis of *wall-associated kinase* (*WAK*) gene family in *Cannabis sativa L. Plants* 11, 2703. doi: 10.3390/plants11202703

Sjölander, K. (2004). Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics* 20, 170–179. doi: 10.1093/bioinformatics/bth021

Stephens, C., Hammond-Kosack, K. E., and Kanyuka, K. (2022). WAKsing plant immunity, waning diseases. *J. Exp. Bot.* 73, 22–37. doi: 10.1093/jxb/erab422

Sun, Z., Song, Y., Chen, D., Zang, Y., Zhang, Q., Yi, Y., et al. (2020). Genome-wide identification, classification, characterization, and expression analysis of the *wall-associated kinase* family during fruit development and under wound stress in tomato (*Solanum lycopersicum* L.). *Genes* 11, 1186. doi: 10.3390/genes11101186

Tamura, K., Stecher, G., and Kumar, S. (2021). MEGA11: Molecular evolutionary genetics analysis version 11. *Mol. Biol. Evol.* 38, 3022–3027. doi: 10.1093/molbev/msab120

Thiltgen, G., Dos Reis, M., and Goldstein, R. A. (2017). Finding direction in the search for selection. *J. Mol. Evol.* 84, 39–50. doi: 10.1007/s00239-016-9765-5

Tian, F., Yang, D.-C., Meng, Y.-Q., Jin, J., and Gao, G. (2019). PlantRegMap: charting functional regulatory maps in plants. *Nucleic Acids Res.* 48, D1104–D1113. doi: 10.1093/nar/gkz1020

Tripathi, R. K., Aguirre, J. A., and Singh, J. (2021). Genome-wide analysis of *wall associated kinase* (*WAK*) gene family in barley. *Genomics* 113, 523–530. doi: 10.1016/j.ygeno.2020.09.045

Verica, J. A., and He, Z.-H. (2002). The *cell wall-associated kinase* (*WAK*) and *WAK-like* kinase gene family. *Plant Physiol.* 129, 455–459. doi: 10.1104/pp.011028

Wagner, T. A., and Kohorn, B. D. (2001). Wall-Associated Kinases are expressed throughout plant development and are required for cell expansion. *Plant Cell* 13, 303–318. doi: 10.1105/tpc.13.2.303

Wang, D., Qin, L., Wu, M., Zou, W., Zang, S., Zhao, Z., et al. (2023). Identification and characterization of *WAK* gene family in *Saccharum* and the negative roles of *ScWAK1* under the pathogen stress. *Int. J. Biol. Macromolecules* 224, 1–19. doi: 10.1016/j.ijbiomac.2022.11.300

Xia, X., Zhang, X., Zhang, Y., Wang, L., An, Q., Tu, Q., et al. (2022). Characterization of the *WAK* gene family reveals genes for FHB resistance in bread wheat (*Triticum aestivum* L.). *Int. J. Mol. Sci.* 23, 7157. doi: 10.3390/ijms23137157

Yan, W., Hu, P., Ni, Y., Zhao, H., Liu, X., Cao, H., et al. (2023). Genome-wide characterization of the *wall-associated kinase-like* (*WAKL*) family in sesame (*Sesamum indicum*) identifies a *SiWAKL6* gene involved in resistance to *Macrophomina phaseolina*. *BMC Plant Biol.* 23, 624. doi: 10.1186/s12870-023-04658-1

Yu, H., Zhang, W., Kang, Y., Fan, Y., Yang, X., Shi, M., et al. (2022). Genome-wide identification and expression analysis of *wall-associated kinase* (*WAK*) gene family in potato (*Solanum tuberosum* L.). *Plant Biotechnol. Rep.* 16, 317–331. doi: 10.1007/s11816-021-00739-5

Zhang, Z., Huo, W., Wang, X., Ren, Z., Zhao, J., Liu, Y., et al. (2023). Origin, evolution, and diversification of the *wall-associated kinase* gene family in plants. *Plant Cell Rep.* 42, 1891–1906. doi: 10.1007/s00299-023-03068-9

Zhang, B., Li, P., Su, T., Li, P., Xin, X., Wang, W., et al. (2020). Comprehensive analysis of *wall-associated kinase* genes and their expression under abiotic and biotic stress in chinese cabbage (*Brassica rapa* ssp. *Pekinensis*). *J. Plant Growth Regul.* 39, 72–86. doi: 10.1007/s00344-019-09964-3

Zhang, Z., Ma, W., Ren, Z., Wang, X., Zhao, J., Pei, X., et al. (2021). Characterization and expression analysis of *wall-associated kinase* (*WAK*) and *WAK-like* family in cotton. *Int. J. Biol. Macromolecules* 187, 867–879. doi: 10.1016/j.ijbiomac.2021.07.163

Zhong, X., Li, J., Yang, L., Wu, X., Xu, H., Hu, T., et al. (2023). Genome-wide identification and expression analysis of *wall-associated kinase* (*WAK*) and *WAK-like* kinase gene family in response to tomato yellow leaf curl virus infection in *Nicotiana benthamiana*. *BMC Plant Biol.* 23, 146. doi: 10.1186/s12870-023-04112-2

Zuo, C., Liu, Y., Guo, Z., Mao, J., Chu, M., and Chen, B. (2019). Genome-wide annotation and expression responses to biotic stresses of the *wall-associated kinase - receptor-like kinase* (*WAK-RLK*) gene family in Apple (*Malus domestica*). *Eur. J. Plant Pathol.* 153, 771–785. doi: 10.1007/s10658-018-1591-8

Chapter 2 - *In silico* characterisation of the avocado *WAK/WAKL* gene family with a focus on genes involved in defence against *Phytophthora cinnamomi*

Independent review at Frontiers in Plant Science – Plant Bioinformatics

Submitted 02 August 2024

In silico characterisation of the avocado *WAK/WAKL* gene family with a focus on genes involved in defence against *Phytophthora cinnamomi*

Aaron Harvey[1,2], Noëlani van den Berg[1,2] and Velushka Swart[1,2*]

[1] Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria, South Africa

[2] Hans Merensky Chair in Avocado Research, Forestry & Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, South Africa

*** Correspondence:**

Velushka Swart - velushka.swart@up.ac.za

Language preference: British English

Keywords: *Persea americana*, Wall-associated kinase, Wall-associated kinase-like, Promoter analysis, Expression analysis, Protein modelling, Protein-ligand binding.

Abstract

The avocado industry faces a significant threat from the hemibiotrophic oomycete pathogen *Phytophthora cinnamomi*. One of the most variably expressed defence genes during an avocado infection trial was a *Wall-associated kinase* (*WAK*). However, further research into this gene family in avocado is limited. WAK and WAK-like (WAKL) proteins are known to bind to fragmented pectin (oligogalacturonides) produced during pathogen penetration, thereby activating downstream defence-related pathways. To better understand the *P. cinnamomi*-avocado defence interaction, this gene family was assessed using *in silico* methods. In this study, previously generated RNA-sequencing data were used to associate genes with the defence response, followed by promoter- and phylogenetic analysis of these genes/proteins. The predicted proteins from these genes were modelled with AlphaFold2, and structural similarity across different rootstocks, as well as their binding affinity for oligogalacturonides, were assessed. The analysis identified 14 *PaWAK*s and 62 *PaWAKL*s across the West-Indian (pure accession reference), Dusa®, Leola™ and R0.12 avocado rootstock genomes. These genes showed distribution across the West-Indian genome's chromosomes, with MCScanX analyses predicting tandem duplications. Differences in gene family composition were

observed across the rootstocks. *PaWAK/WAKL* expression profiles were compared between *P. cinnamomi*-infected R0.12 (susceptible) and Dusa® (partially resistant) rootstocks, implicating five *PaWAK/WAKL*s in defence. Phylogenetic and promoter analyses were conducted to predict associated defence-related pathways, focusing on stress and phytohormone-responsive pathways. It is hypothesized that these pathways influence the regulation of the upregulated *PaWAK/WAKL*s. The proteins of the upregulated genes showed structural differences and varying oligogalacturonide binding affinities across rootstocks. This is the first comprehensive *in silico* characterisation of the *PaWAK/WAKL* gene family in avocado. The analyses revealed differences across rootstocks, indicating that *PaWAK/WAKL*s play a crucial role in the varying efficiencies of defence responses at a multi-omic level. These genes could be incorporated into a molecular screening tool to improve the development of resistant avocado rootstocks.

Introduction

Persea americana (avocado) production is significantly threatened by the soil-borne oomycete *Phytophthora cinnamomi* (Hardham, 2005, Hardham and Blackman, 2018). Infection by this hemibiotroph can lead to Phytophthora root rot, a disease that results in decreased fruit yield and potential tree death (Ramírez-Gil *et al.*, 2017). Avocado deploys a multi-faceted defence response that includes cell wall modifications, Reactive Oxygen Species (ROS) scavenging and detoxification, proteinase inhibitors and induction of defence-related genes (van den Berg *et al.*, 2021). A *Wall-associated kinase* (WAK) was identified as one of the defence- and stress-related genes with the most variable expression between the susceptible R0.12 and partially-resistant Dusa® avocado rootstocks, following *P. cinnamomi* inoculation (Backer *et al.*, 2022). The *WAK* and *WAK-like* (*WAKL*) gene family has been associated with defence responses in a variety of plant species. Examples that illustrate the extensive defence capabilities of *WAK/WAKLs* in defence against bacterial, fungal, viral and oomycete pathogens include *WAK/WAKL*s in tomato against *Pseudomonas syringae*, rose against *Botrytis cinerea*, cotton against *Verticillium dahliae* and *Fusarium oxysporum*, rice against *Magnaporthe oryzae, Nicotiana benthamiana* against tomato yellow leaf curl virus, and potato against *Phytophthora infestans* (Delteil *et al.*, 2016, Wang *et al.*, 2020, Zhang *et al.*, 2020b, Liu *et al.*, 2021, Yu *et al.*, 2022, Zhong *et al.*, 2023). Some WAK/WAKLs have been shown to confer

resistance to fungi with either hemibiotrophic or necrotrophic lifestyles, through various mechanisms including pathogen- or host-derived elicitor detection and cell wall restructuring (Stephens *et al.*, 2022).

WAK proteins are transmembrane pectin receptors, part of the 15 receptor-like kinase (RLK) subfamilies, serving several functions within the plant cell. WAKs can bind native pectin thereby influencing cell expansion during developmental stages. They can also bind oligogalacturonides (OGs), which are fragmented pectin produced during abiotic stress like wounding and biotic stress like pathogen infection. The OGs act as damage-associated molecular patterns (DAMPs) to elicit a defence response (Kohorn and Kohorn, 2012, Kohorn, 2015).

With the availability of the avocado West-Indian pure-accession reference genome and the re-sequenced genomes of key rootstocks (Avocado Genome Consortium), there is an opportunity to identify and compare the *PaWAK/WAKLs* among rootstocks with varying levels of resistance to *P. cinnamomi*. Given the importance of this gene family in plant defence across multiple species, it is hypothesised that a subset of these genes will be implicated in the successful defence response of avocado against *P. cinnamomi*. Additionally, differences in this gene family between rootstocks are thought to contribute to variation in defence. In total, 14 *Persea americana PaWAK*s and 62 *PaWAKL*s were identified and fully described. Five candidates were implicated in defence, showing differences between the rootstocks in terms of expression during infection, *cis*-acting regulatory elements, predicted 3D protein structures and OG binding affinity (BA).

Materials and methods

Putative *PaWAK/WAKL* identification

Wall-associated kinases (WAK) or Wall-associated kinase-like (WAKL) protein sequences were extracted from the NCBI database, including those of AtWAK/WAKLs (https://www.ncbi.nlm.nih.gov/website, accessed on 15/06/2023). A Clustal Omega alignment (Geneious Prime v2023.1.2 with default settings) was performed to generate a HMMER profile, which was used as a query for a local HMMER v3.3.2 (Finn *et al.*, 2011) search against the avocado West-Indian pure accession reference, Dusa®, Leola™ and R0.12

predicted proteomes (Avocado Genome consortium, unpublished data). Results with E-values < $1x10^{-5}$ were retained for further classification.

PaWAK/WAKL classification

All protein sequences were extracted from the West-Indian genome for protein domain analysis. Each sequence had to contain either a WAK or GUB_WAK_bind (Wall-associated receptor kinase galacturonan binding), STKc_IPAK (Catalytic domain of the Serine/Threonine kinases, Interleukin-1 Receptor Associated Kinases and related STKs) or PKc-like (Protein Kinases, catalytic domain) domain predicted by the NCBI Conserved Domain Database (CDD) search (https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi). Additionally, each sequence required a GUB_WAK_Bind and protein kinase (pkinase) domain predicted by InterProScan (specifically the Pfam database https://www.ebi.ac.uk/interpro/) to be considered a WAK/WAKL. Sequences not meeting these criteria were excluded. Subsequent WAK classifications required a WAK or GUB_WAK and STKc_IPAK or PKc-like domain, a signal peptide predicted through SignalIP-5.0 (https://services.healthtech.dtu.dk/services/SignalP-5.0/), a transmembrane domain predicted through THMM-2.0 (https://services.healthtech.dtu.dk/services/TMHMM-2.0/), and at least one EGF (epidermal growth factor)-like or EGF_CA (EGF-calcium binding) domain predicted through NCBI CDD, InterProScan (Pfam), and SMART (http://smart.embl-heidelberg.de/). Domains had to be present in two of three predictions with an E-value < 0.01 to be considered true domains. Sequences not containing these five domains in tandem were classified as WAKLs (Zhang *et al.*, 2021, Li *et al.*, 2022). The PaWAK/WAKLs were renamed according to position on the chromosomes. The protein domains, gene structure and position on the West-Indian genome of the *PaWAK/WAKL* genes were visualised with TBtools v2.086 (Chen *et al.*, 2020).

PaWAK/WAKL protein sequence analyses

The molecular weight and isoelectric point of the proteins, along with their subcellular localization and conserved motifs, were predicted using the online platforms GeneInfinity (http://www.geneinfinity.org/sms/sms_proteinmw.html,), WoLF PSORT (https://wolfpsort.hgc.jp/,) and MEME (https://meme-suite.org/meme/tools/meme). The protein sequences of all the PaWAK/WAKLs were subjected to an All-to-All BLASTp search, followed by the use of MXScanX (in TBtools) to predict the tandemly duplicated gene pairs (Wang *et al.*, 2012, Chen *et al.*, 2020). This analysis was further refined by predicting the non-

synonymous substitution rate (K_a), synonymous substitution rate (K_s), and K_a/K_s ratio for the tandemly duplicated pairs using TBtools v2.086. These predictions were used to determine which evolutionary processes, either purifying or positive selection, are acting on these gene pairs (Chen *et al.*, 2020).

PaWAK/WAKL expression analysis during *P. cinnamomi* infection

RNA-sequencing data were used from a previous study (Backer *et al.*, 2022). In this study, avocado rootstocks were inoculated with a *P. cinnamomi* zoospore suspension, and roots were harvested at 6, 12, 24 and 120 hpi for RNA-sequencing. The expression profiles for all *PaWAK/WAKLs* at these four time points were obtained and visualized through a heatmap generated with R-studio and TBtools v2.086 (Racine, 2012, Chen *et al.*, 2020). Significant expression changes were defined by an adjusted P-value < 0.05 and $Log_2FC > 1$.

Phylogenetic analysis of PaWAK/WAKLs

Arabidopsis thaliana WAK/WAKL protein sequences were obtained from The Arabidopsis Information Resource (TAIR) database (https://www.arabidopsis.org/) based on the designations previously described (Verica and He, 2002). AtWAKL19 was excluded as its sequence was not available on TAIR. Defence-related WAK and WAKL protein sequences were obtained from the NCBI database (**Table S1**, accessed on 11/07/2023) based on relevant literature (Larkan *et al.*, 2020, Wang *et al.*, 2020, Zhang *et al.*, 2020b, Liu *et al.*, 2021, Stephens *et al.*, 2022, Kong *et al.*, 2023). These sequences were aligned with the PaWAK/WAKLs using Clustal Omega, and identical sites were masked. A Jukes-Cantor, Unweighted Pair-Group Method with Arithmetic Mean (UPGMA) phylogenetic tree was constructed using a bootstrap of 5000 (node threshold of 50%), with default settings on Geneious Prime v2023.1.2.

Promoter analysis of *PaWAK/WAKLs*

The promoter regions (2000 bp upstream of the start codon) of all identified *PaWAK* and *PaWAKL* genes were extracted from the West-Indian pure accession, Dusa®, Leola™, and R0.12 genomes using GenomeView-N42 (Abeel *et al.*, 2012, Zhang *et al.*, 2021, Li *et al.*, 2022, Wang *et al.*, 2023, Zhong *et al.*, 2023). These promoter sequences were uploaded to the PlantRegMap Binding site prediction tool (http://plantregmap.gao-lab.org/binding_site_prediction.php on 04/08/223) with *A. thaliana* as the reference species and a P-value cutoff of < 10^{-5} (Lescot *et al.*, 2002, Tian *et al.*, 2019). Additionally, the promoter sequences were analysed using the PlantCARE online database

(https://bioinformatics.psb.ugent.be/webtools/plantcare/html/, accessed on 26/09/2023). Identified elements with two or more predictions in the same location were considered as a single element. The identified elements were visualized with TBtools v2.086.

Structure and ligand binding analyses of candidate defence PaWAK/WAKLs

The protein sequences corresponding to the upregulated genes were uploaded to the AlphaFold2 online platform for 3D structure prediction (https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb#scrollTo=kOblAo-xetgx). AlphaFold2 was chosen because AlphaFold3 was still in Beta version at the time of article submission. The resulting protein structure PDB files were visualized with ChimeraX v1.7.1, and the protein domains were coloured according to the InterProScan Pfam protein domain predictions.

The 3D protein structures of the same WAK/WAKL proteins across different genomes were superimposed using CLICK and TM-score server (http://cospi.iiserpune.ac.in/click/ and https://seq2fun.dcmb.med.umich.edu//TM-score/ - accessed on 28/02/2024) to calculate superimposition scores between the pairs (Zhang and Skolnick, 2004, Nguyen *et al.*, 2011). Through the CLICK analysis, lower to middling RMSD scores (< 2 Å to 2-4 Å) were used to indicate high to moderate structural similarity, while higher to middling associated Z-scores (> 2 and 1-2) were used to suggest significant alignment better than random chance. A high RMSD score (> 2 Å) with a low Z-score (< 1) indicated structural differences and unreliable alignment. The RMSD scores followed the same boundaries in the TM-score sever analysis, with scores closer to one (cutoff boundary TM-score > 0.5), indicating that the proteins are more likely to be in the same fold (Xu and Zhang, 2010). To serve as negative and positive controls, different PaWAKLs from two genomes (PaWAKL17 from Dusa® and PaWAKL61 from Leola™) and the same PaWAKL from the same genome (PaWAKL17 from Dusa®) were superimposed.

The 3D protein structures of the PaWAK/WAKLs were uploaded to the Molecular Docking server platform (https://www.dockingserver.com/, accessed on 01/03/2024) along with an OG 3D structure (beta-D-Galactopyranuronic acid, pectin monomer from PubChem with accession number 441476: https://www.ncbi.nlm.nih.gov/, accessed on 19/02/2024) to predict binding energies and inhibition concentration (K_i) of the OG interaction with the PaWAK/WAKLs (Bikadi and Hazai, 2009). Because this interaction did not involve an inhibitor

as the ligand, K_i was interpreted as K_d (dissociation constant), reflecting the BA of the protein for the ligand. Free energy of binding quantifies the energy change or stability that occurs due to the binding taking place. Favourable ligand-protein binding is indicated as negative free energy of binding values, with the more negative values (lower values) indicating greater favourability. The vdW + Hbond + desolv energy is a single value that quantifies the energy contributed by van der Waals forces, hydrogen bonds, and desolvation energy, with larger negative values representing more stable/favourable interactions. The K_d represents the concentration at which half of the protein's binding sites are occupied with the ligand, with lower K_d values suggesting tighter binding affinity and higher values representing weaker binding affinity. The target binding area on the PaWAK/WAKLs was centred at the GUB_WAK_bind domain. The target box started with a volume of 35 $Å^3$ and was reduced to ensure the GUB_WAK_bind domain and minimal other amino acids were included. The best binding scenario was used for evaluation.

Results

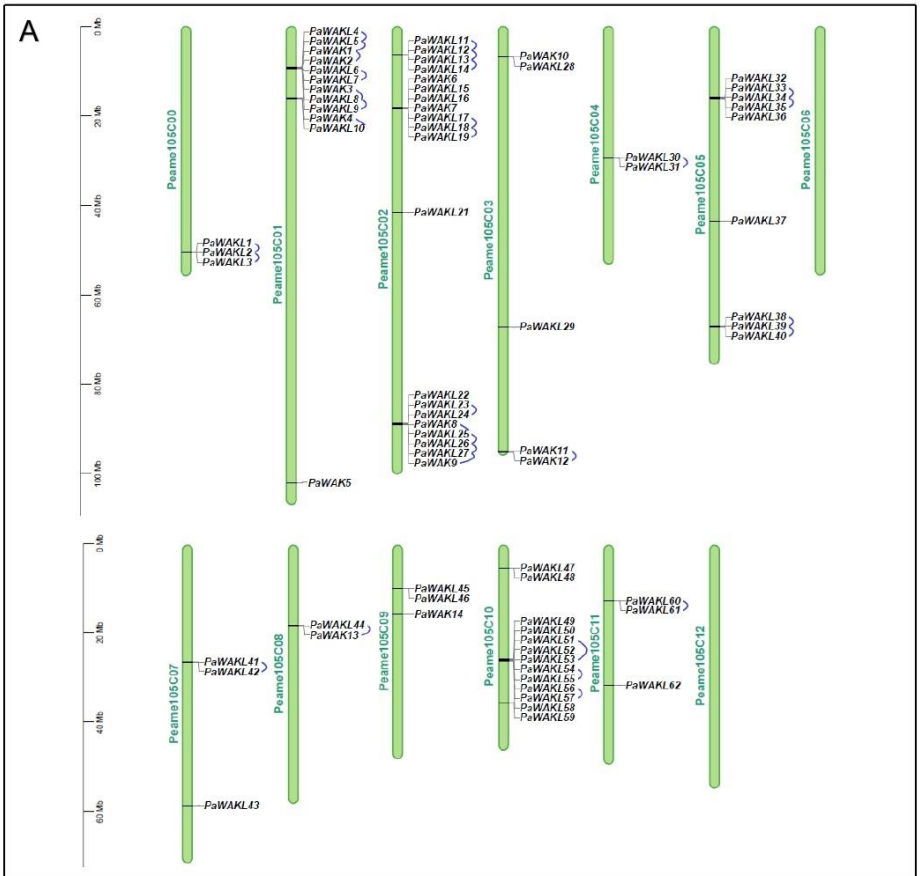
The full repertoire of *PaWAK* and *PaWAKL* genes characterised

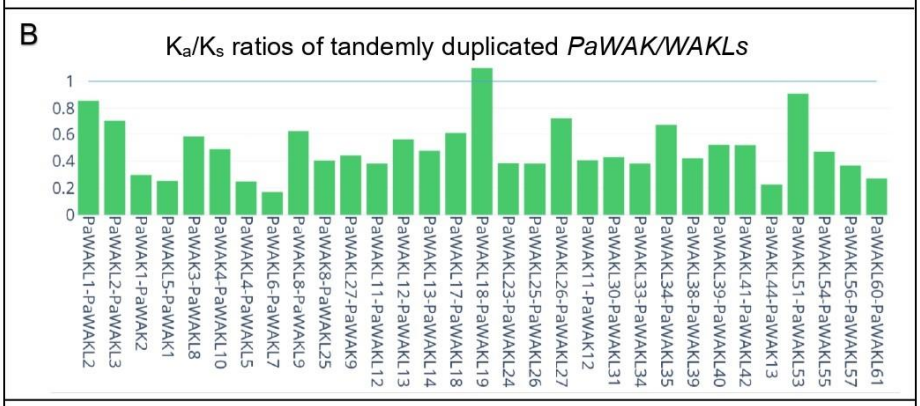
The first step in the characterisation of the *PaWAK/WAKL*s was to identify, classify and describe the gene family members. The West-Indian reference genome contained 14 *PaWAKs* and 62 *PaWAKLs* that were randomly distributed, with some clustering across 10 of the 12 chromosomes and the unassigned scaffolds (**Figure 1A**), similar to observations in other species (Zhang *et al.*, 2020a, Liu *et al.*, 2021, Yu *et al.*, 2022, Li *et al.*, 2023, Yan *et al.*, 2023). There were 31 tandemly duplicated pairs predicted, with K_a/K_s ratios between 0.17 and 1.10 (**Figure 1B** and **Table S2**). One pair (*PaWAKL18* and *PaWAKL19*) had a ratio > 1, suggesting positive selection for this pair, unlike the rest of the gene pairs predicted to be undergoing purifying selection ($K_a/K_s < 1$).

When comparing the genomes of R0.12, Leola™ and Dusa® to the West-Indian reference, all 14 *PaWAKs* were identified in the other three genomes. However, the number of *PaWAKL*s varied: R0.12 had 60, Leola™ had 56 and Dusa® had 58, with no additional *PaWAK/WAKLs* being identified (**Table S3**). Dusa® lacked *PaWAKL3*, *PaWAK6*, *PaWAK16*, and *PaWAK32*; R0.12 lacked *PaWAKL16* and *PaWAK39*; Leola™ lacked *PaWAKL12*, *PaWAK14*, *PaWAK15*, *PaWAK16*, *PaWAK21*, and *PaWAK53*. *PaWAKL16* was unique to the West-Indian genome

while *PaWAKL39* was only absent in R0.12 (a *P. cinnamomi*-susceptible rootstock) but present in both partially-resistant rootstocks (Leola™ and Dusa®).

The *PaWAK/WAKL* open reading frames (ORFs) varied in length from 2,000 to 30,000 nucleotides, comprised of 1-10 exons and introns of differing lengths (**Figure 1C**). The gene architectures fell within the range of the previously described smallest (1.3 kb in *N. benthamiana*) and largest (~330 kb in *Juglans* species) *WAK/WAKLs*, with exon numbers showing similarities to other characterized *WAK/WAKLs*, such as the widest range of 1-27 exons in *Brassica rapa* (Zhang *et al.*, 2020a, Li *et al.*, 2022, Zhong *et al.*, 2023). In *A. thaliana* (*At*)*WAK/WAKL*s, the gene structures were well conserved, with three exons and two introns, where the middle exon was smaller than the two flanking it (Verica and He, 2002). This gene structural pattern was seen in 11/14 *PaWAK*s and 29/62 *PaWAKL*s, representing the majority (40/62) of the gene structures observed. This pattern, including those with four exons and three introns, is associated with immunity-related *WAK* genes in species like *A. thaliana,* cotton and rice (Stephens *et al.*, 2022).





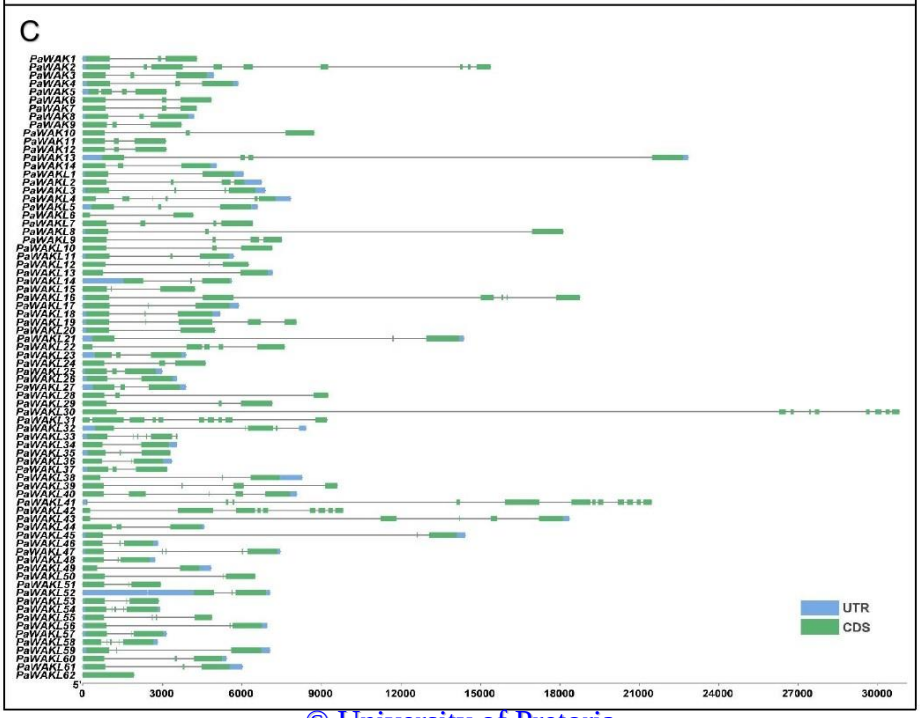


Figure 1: *PaWAK/WAKL* gene family characteristics. (**A**) The *PaWAK/WAKL*s are distributed across 12 chromosomes and unassigned scaffolds (Peame105C00) of the West-Indian pure accession reference genome, with tandemly duplication pairs connected by blue lines. (**B**) The K_a/K_s ratio for the tandemly duplicated pairs was calculated using MCScanX. Pairs with a K_a/K_s ratio greater than one were suggested to be undergoing positive selection, while ratios below one indicated purifying selection. (**C**) The open reading frames of the *PaWAK/WAKL*s ranged from approximately 2000 to 30 000 bp, with varying numbers of exons (coding sequences - CDS) and introns (grey lines), and some untranslated regions (UTRs) identified. Chromosome placements and gene structures were visualized with TBtools, and the K_a/K_s values were visualized with Plotly (https://chart-studio.plotly.com).

PaWAK/WAKL protein properties and conserved motifs

The downstream PaWAK/WAKL proteins were also characterized to provide a multi-level description of this family. The proteins ranged from 301 to 1293 amino acids in length, had molecular weights between 33.45 and 142.86 kD, and isoelectric points between 5.17 and 9.13. The majority were predicted to localize to the plasma membrane, with some found throughout the cell (**Figure S1** and **Table S4**). These properties were within normal ranges for plant species, and plasma membrane localisation has been experimentally demonstrated in cotton and bread wheat WAK/WAKLs (Ramírez-Sánchez *et al.*, 2016, Mohanta *et al.*, 2019, Zhang *et al.*, 2021, Xia *et al.*, 2022). All PaWAK proteins contained the five required domains, while PaWAKLs displayed multiple domain combinations, showing variation in the presence or absence of EGF_like, TM and SP domains. Specifically, 14/62 WAKLs lacked a signal peptide and 26/62 lacked a transmembrane domain (**Figure 2A** and **Table S3**).

PaWAKL5 was predicted to have an EGF_CA domain in the NCBI Conserved Domain Database (CDD) search with an E-value of ~0.002. However, this domain was not predicted through the Pfam database on InterProScan, and the SMART search E-value for the domain was 0.028 (E-value > 0.01, not considered significant), thus it was classified as a WAKL. The 10 most abundant predicted conserved motifs in the PaWAK/WAKL proteins clustered within the C-terminal region, where the kinase domain lies (**Figure 2B**). When comparing the pairwise identity of the same WAK/WAKL protein sequences across the four genomes, identity varied from 10.5% to 100%, with more than half (47/76) showing an identity above 80%. The protein sequence variation was mostly present in PaWAKLs, while PaWAK sequences were more

conserved. This could suggest differences in the proteins produced by the same gene from different rootstocks.

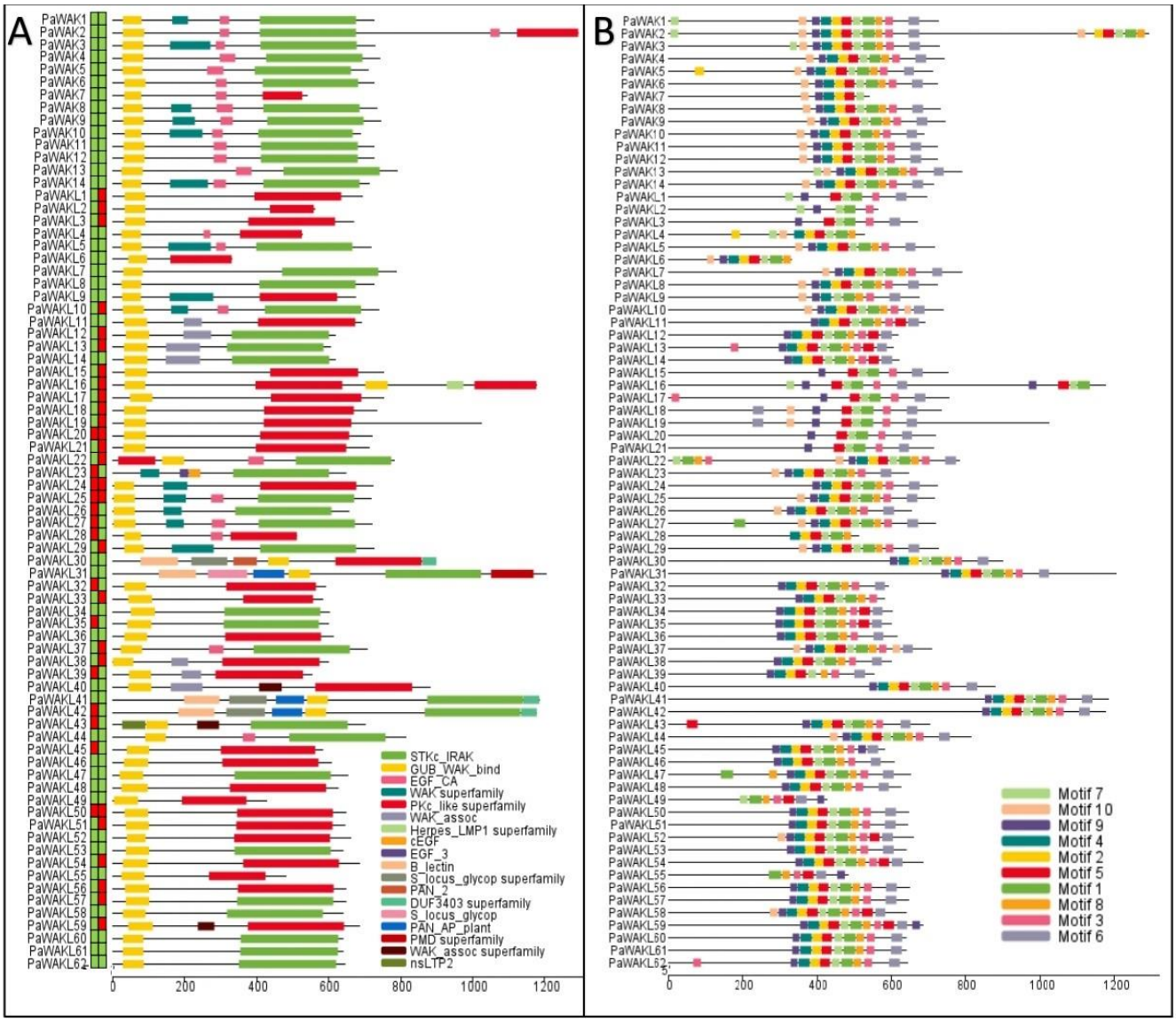


Figure 2: PaWAK/WAKL protein domains and conserved motifs. (**A**) PaWAK and PaWAKL protein domains were visualised using the results from the NCBI Conserved Domain Database (CDD) batch search. SignalP-5.0 and TMHMM-2.0 were used to predict the presence of signal peptides and transmembrane domains, with the blocks preceding the proteins representing the presence (green) or absence (red) of each, respectively in that order. (**B**) The top 10 most conserved protein motifs were predicted through MEME. Protein properties were visualised with Plotly (https://chart-studio.plotly.com), and protein domains and conserved motifs were visualized with TBtools.

Expression analysis of *PaWAK/WAKLs* during *P. cinnamomi* infection

The upregulation of certain genes during infection strongly implicated them in pathogen defence responses. The *PaWAK/WAKL* expression profiles, from RNA-sequencing data (avocado inoculated with *P. cinnamomi*), showed differential expression within this gene family (**Figure 3** and **Table S5**). Significant upregulation at one or more time points was observed for 22 (28.9%) *PaWAK/WAKLs* in Dusa® and for 14 (18.4%) *PaWAK/WAKLs* in R0.12. In the partially-resistant avocado rootstock, Dusa®, *PaWAK4*, *PaWAK13*, *PaWAKL5*, *PaWAKL10*, *PaWAKL29*, and *PaWAKL31* were upregulated at multiple time points. Notably, *PaWAK13* and *PaWAKL29* also showed upregulation at 6 hours post-inoculation (hpi) in the susceptible R0.12. *PaWAKL29* exhibited a Log_2Fold Change (Log_2FC) > 8 at three time points (6, 12, 24 hpi) in Dusa®. In contrast, in R0.12, *PaWAKL29* expression was > 7 Log_2FC at 6 hpi, reducing to a non-significant increase compared to the experimental control at subsequent time points. Significant downregulation at one or more time points was seen for 10 (13.2%) *PaWAK/WAKLs* in Dusa® and for nine (11.8%) *PaWAK/WAKLs* in R0.12. Downregulation across two or more time points in Dusa® was observed for *PaWAKL17, PaWAKL18, PaWAKL43, PaWAKL49,* and *PaWAKL61.* Of these, only *PaWAKL61* did not show downregulation in R0.12 as well.

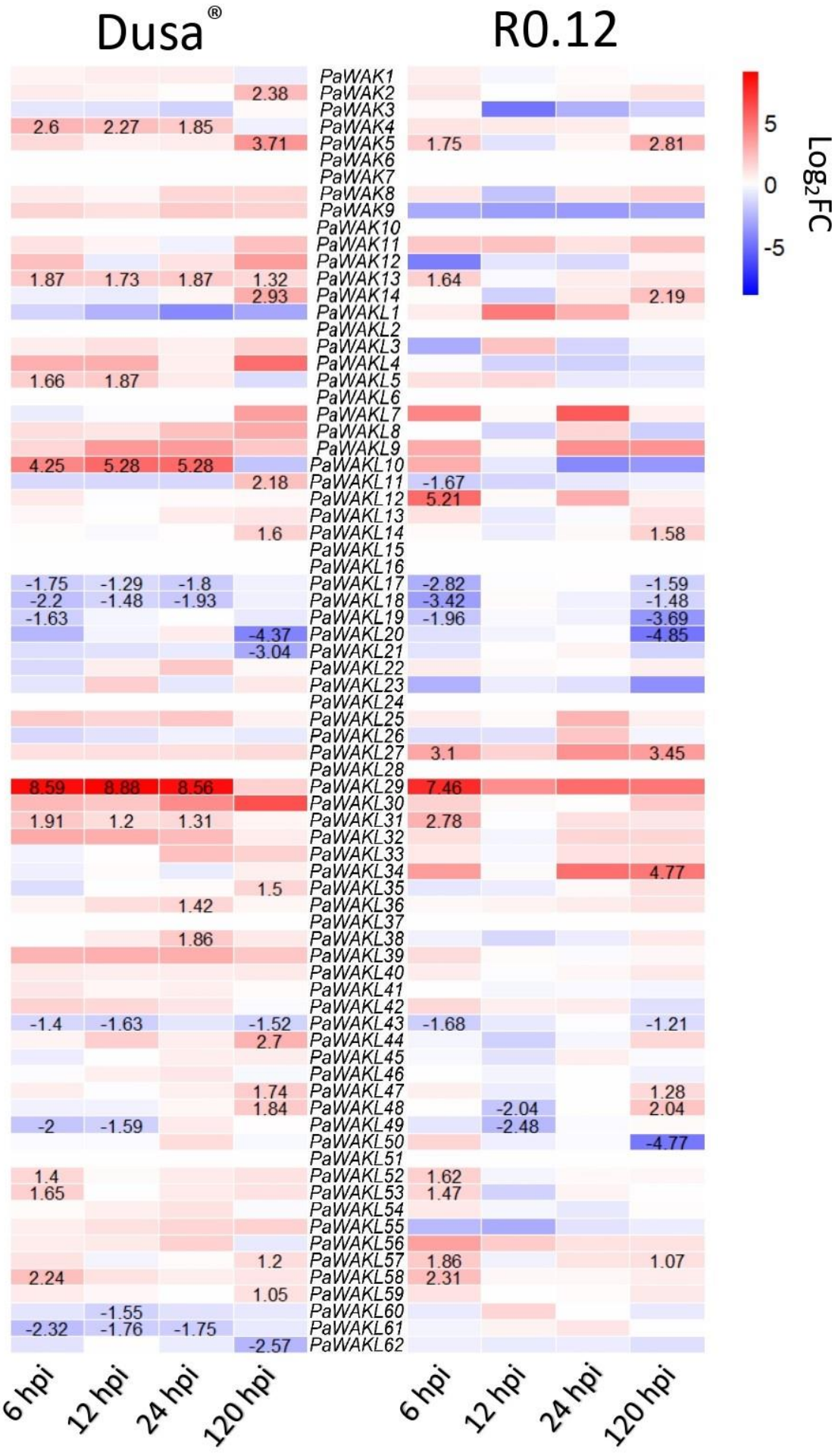


Figure 3: Expression profile heatmap of all *PaWAK/WAKL* genes during *P. cinnamomi* infection of a partially-resistant (Dusa®) and a susceptible (R0.12) avocado rootstock. More intense red blocks indicate a higher Log_2FoldChange (Log_2FC) compared to the uninoculated controls and more intense blue blocks indicate a lower Log_2FC compared to the controls. Values are shown for time points where the Log_2FC change was significant, defined as adjusted *P*-value < .05 and Log_2FC > 1. All Log_2FC values and adjusted *P*-value are provided in **Table S5.**

Phylogenetic analysis of PaWAK/WAKL, AtWAK/WAKL and defence-related WAK/WAKLs

The Jukes-Cantor Unweighted Pair-Group Method with Arithmetic Averaging (UPGMA) phylogenetic tree was constructed to determine if any PaWAK/WAKL showed a relationship to previously characterized proteins, allowing for functional inferences (**Table S1**). The analysis revealed that the majority of the WAK/WAKLs from other species clustered separately from the PaWAK/WAKLs (**Figure 4**). Most of the AtWAK/WAKLs formed distinct clusters, some of which included WAK/WAKLs from other species. However, AtWAKL7, AtWAKL14, AtWAKL15, AtWAKL20, AtWAKL 21, and most of the defence-related WAK/WAKLs included in the analysis (excluding Rlm9 and RcWAK4) were dispersed within the PaWAK/WAKL clusters.

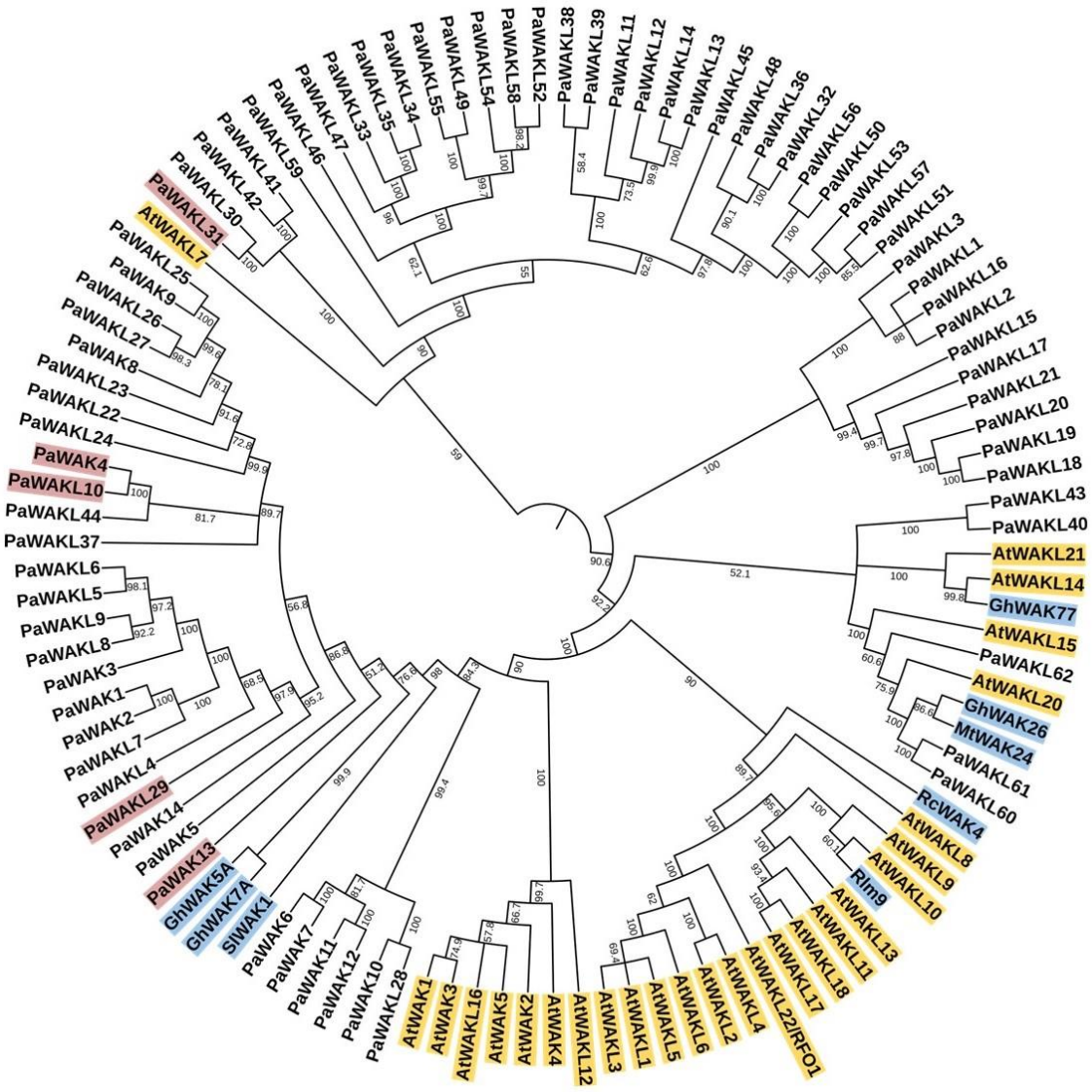


Figure 4: Phylogenetic analysis of the PaWAK/WAKL protein sequences with WAK/WAKLs from *Arabidopsis thaliana* and a subset associated with defence in rose, cotton, tomato, and barrel medic. This represents a Jukes-Cantor, Unweighted Pair-Group Method with Arithmetic Averaging (UPGMA) phylogenetic analysis of PaWAK/WAKL, AtWAK/WAKL (yellow) protein sequences along with those implicated in defence from other species in blue (Larkan *et al.*, 2020, Zhang *et al.*, 2020b, Liu *et al.*, 2021, Stephens *et al.*, 2022, Kong *et al.*, 2023). Bootstrap support values are shown preceding the nodes (rounded to one decimal place). Protein sequences highlighted in red correspond to genes that showed upregulation during *P. cinnamomi* infection. The tree was constructed using Geneious v2023.1.2 with a bootstrap of 5000, using default settings. The Newick tree was visualized with Interactive Tree of Life (iTOL v6, https://itol.embl.de/), and the protein sequences are provided in **Table S1.**

The protein sequences corresponding to the *PaWAKs* and *PaWAKLs* that are located near each other on the genome were grouped together in the phylogenetic analysis. This suggested that sequences in close genomic proximity tend to be more similar than those further apart. This correlated with the high number of tandemly duplicated gene pairs, as this expansion allows for duplication near the original sequence. An example was the clustering of *PaWAK9, PaWAKL25*, *PaWAKL26,* and *PaWAKL27*, which were all located in proximity to one another on the chromosome, with the *PaWAKL25* and *PaWAKL26* pair, and the *PaWAKL26* and *PaWAKL27* pair identified as tandemly duplicated genes. The phylogenetic analysis supported the evidence of tandem duplications within this cluster. The node bootstrap confidence values ranged from 51.2 to 100, with the majority being over 70 (88/102 nodes > 70), allowing for functional inferences – sequences closely related may have similar functions within the plant cell. The closest relationships between a PaWAK/WAKL protein identified during the expression analyses, and defence-related proteins from other plant species were between PaWAK13 and *Gossypium hirsutum* (cotton) GhWAK5A, GhWAK7A and *Solanum lycopersicum* (tomato) SlWAK1 (Wang *et al.*, 2020, Zhang *et al.*, 2020b). Pairwise comparisons of the protein sequences of PaWAK13 and SlWAK1, GhWAK5A and GhWAK7A, showed percent identities of 42.2%, 44.8% and 42.2%, respectively.

Prediction of *cis*-acting elements for inferring involvement in cellular pathways

To understand the factors influencing the differential expression between Dusa® and R0.12 during infection, the promoter regions (2000 bp upstream of the start codon) of the upregulated genes showing upregulation across three or more timepoints in Dusa® (*PaWAK4*, *PaWAK13*, *PaWAKL10*, *PaWAKL29*, and *PaWAKL31*) were analysed for the presence and composition of *cis*-acting elements (**Figure 5**). The low number of identified elements in the promoter of *PaWAKL29* in R0.12 was due to a large portion of the promoter consisting of unannotated nucleotides (N), making it difficult to draw conclusions on the *PaWAKL29* regulation in R0.12. While the focus was on the elements identified in the promoters of upregulated genes, the full repertoire for all *PaWAK/WAKLs* was predicted (**Table S6**). In total, 74 unique elements were identified with roles in light- (23 elements), phytohormone- (10 elements), and stress-responsiveness (three elements), as well as general processes (five elements) and floral development (three elements). Certain elements such as ACE, AE-box, ATCT-motif, AuxRR-core, Box II, Circadian, GA-motif, Gap-box, GTGGC-motif, LTR, Pc-CMA2c

TGA-element, ARR-B, BBR-BPC, BES1, EF2/DP, GeBP, LYF, Nin-like, and SRS showed presence in only one gene (gene specificity) when considering the promoter regions of the five upregulated genes.

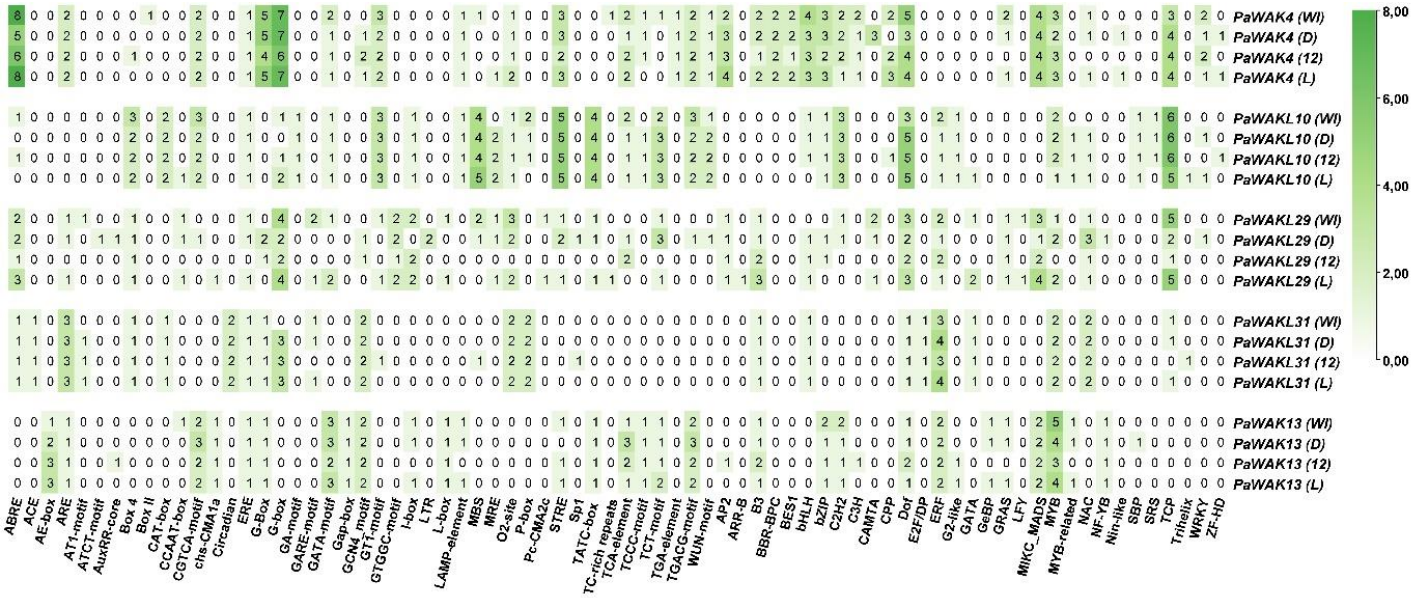


Figure 5: *Cis*-acting element composition in the promoter regions (2000 bp upstream of the start codon) of *PaWAK4, PaWAKL10, PaWAKL29, PaWAKL31, and PaWAK13* across four avocado rootstock genomes (WI-West Indian, D-Dusa®, 12-R0.12 and L-Leola™) as predicted by PlantCARE (ABRE to TGACG-element) and PlantRegMap (AP2 to ZF-HD). More intense green indicates a greater number of unique locations for that element, with the value representing the number of unique locations predicted.

Elements found in high abundance (three or more) in the *PaWAK4* promoter regions across the genomes included ABRE, G-Box, G-box, STRE, AP2, bHLH, bZIP, Dof, MICK_MADS, MYB and TCP. These elements are involved in phytohormone responsiveness, growth/development, and environmental responses (Tian *et al.*, 2019). Notably, ABRE, an abscisic acid-responsive element, was more abundant in *PaWAK4* than in the other four genes, potentially indicating a pathway that differentiates its expression from the others. *PaWAKL10* had fewer elements overall, compared to *PaWAK4*, with the same high abundance of Dof and TCP. However, *PaWAK4* lacked the TATC-box (involved in gibberellin-responsiveness), while *PaWAKL10* contained four copies, conserved across the promoters from all four genomes. *PaWAKL29*, *PaWAKL31* and *PaWAK13* had fewer elements overall in the upregulated subset of promoters.

When comparing *PaWAK/PaWAKL* promoters in Dusa® to those in R0.12, there were 11 (*PaWAK4*), 1 (*PaWKL10*), 25 (*PaWAKL29*), 3 (*PaWAKL31*), and 11 (*PaWAK13*) elements that were more abundant in Dusa®. Alternatively, there were 8 (*PaWAK4*), 4 (*PaWKL10*), 5 (*PaWAKL29*), 4 (*PaWAKL31*), and 6 (*PaWAK13*) that were in higher abundance in R0.12 promoters when compared to Dusa®.

The promoter regions of an additional partially-resistant avocado rootstock, Leola™, were included to predict whether the pathways involved in the upregulation of *PaWAK4*, *PaWAK13*, *PaWAKL10*, *PaWAKL29*, and *PaWAKL31* in Dusa®, were similar in Leola™. The Leola™ *PaWAK4* promoter contained additional ABRE and STRE elements compared to R0.12, and additional ABRE elements compared to Dusa®, indicating a potentially higher responsiveness to abscisic acid (ABA) and stress-related pathways. The Leola™ *PaWAKL10* promoter had a CCAAT-box (stress-responsive element) that was absent in the promoters from all other genomes. Similar to the Dusa® *PaWAKL31* promoter, the Leola™ promoter had a GARE-motif (Gibberellin responsive element), which was absent in R0.12. The *PaWAK13* promoter in Leola™ did not contain any phytohormone- or stress-responsive elements in higher abundance compared to R0.12, but it did have more GATA-motif, I-box, Lamp-element, and TCT-motif elements, which are all light-responsive. This suggested that *PaWAK13* in Leola™ may be more responsive to changes in light, potentially relating to circadian pathways.

Structure and ligand binding analyses of candidate defence PaWAK/WAKLs

The monomer 3D protein structures of PaWAK4, PaWAK13, PaWAKL10, and PaWAKL29 from the four available rootstocks were computationally predicted to assess structural similarity. PaWAKL31 was excluded from this analysis as it could not be modelled with AlphaFold2, for any of the available genomes. Visual representation of the protein structures revealed a lack of structural consistency across rootstocks for the same PaWAK/WAKL proteins (**Figure 6A**). Amongst the proteins analysed, PaWAK13 showed the most similarity, while PaWAKL29 was the least similar. When superimposed, the same WAK or WAKL proteins from all four rootstocks demonstrated even more evident structural discrepancies (**Figure 6B**). Complete superimposition was not achieved for any of the four PaWAK/WAKLs across the rootstocks, only highlighting qualitative structural differences.

Pairs of 3D protein structures from two rootstocks at a time (an example would be R0.12 PaWAK13 superimposed with West-Indian PaWAK13) were assessed to determine if the proteins were structurally identical across the four rootstocks (**Figure 6C** and **Table S7**). Most pairs of superimposed PaWAK4, PaWAK13 and PaWAKL10 had root-mean-square deviation (RMSD) values < 2 Å (Armstrong) with associated Z-scores > 2 and TM-scores > 0.5, indicating high structural similarities between these proteins across the rootstocks. However, PaWAKL29 proteins showed the most variation with only PaWAKL29 from Leola™ and the West Indian being predicted as structurally similar. When compared to R0.12 both PaWAK4 and PaWAK10 in Dusa®, and PaWAK4 in Leola™ showed TM-scores < 0.5, but all other values indicated structural similarity. When compared to the West-Indian PaWAK13 from Dusa®, PaWAK10 from R0.12, and PaWAKL29 from Leola™ had RMSD values > 2 Å from the TM-score server, although the CLICK server RMSD scores were < 2 Å with high Z-scores and TM-scores, indicating structural similarity. The comparisons of PaWAKL10 in Leola™ and West Indian, PaWAKL29 in Dusa® and West Indian and PaWAKL29 in Dusa® and Leola™ had RMSD values > 2 Å from the TM-score server, with associated low TM-scores, but a low RMSD from CLICK, and high Z-scores. The low Z- and TM-scores for PaWAKL29 in R0.12 and West Indian, PaWAKL29 in Dusa® and R0.12, and PaWAKL29 in Leola™ and R0.12 suggest that these pairs consist of structurally different proteins.

A PaWAK/WAKL-OG binding prediction analysis was conducted to assess the binding potential of each corresponding upregulated gene's PaWAK/WAKL protein with an OG and to identify if any proteins show higher affinity for the OG ligand, given the structural differences between rootstocks. All the PaWAK/WAKLs assessed had negative binding energies between -4 and -5 kcal/mol, indicating favourable exothermic (spontaneous) binding (**Figure 6D** and **Table S8**). There were binding energy and dissociation constant (K_d) differences between the same PaWAK/WAKL proteins from different rootstocks, likely due to the structural differences noted previously. The proteins that were predicted to bind OGs most efficiently from each rootstock were PaWAK4 from R0.12, PaWAK13 from Leola™, PaWAKL10 from West-Indian and PaWAKL29 from R0.12. However, the Dusa® and Leola™ versions of PaWAK13 and PaWAKL10 were predicted to bind OGs more efficiently than their R0.12 equivalents.

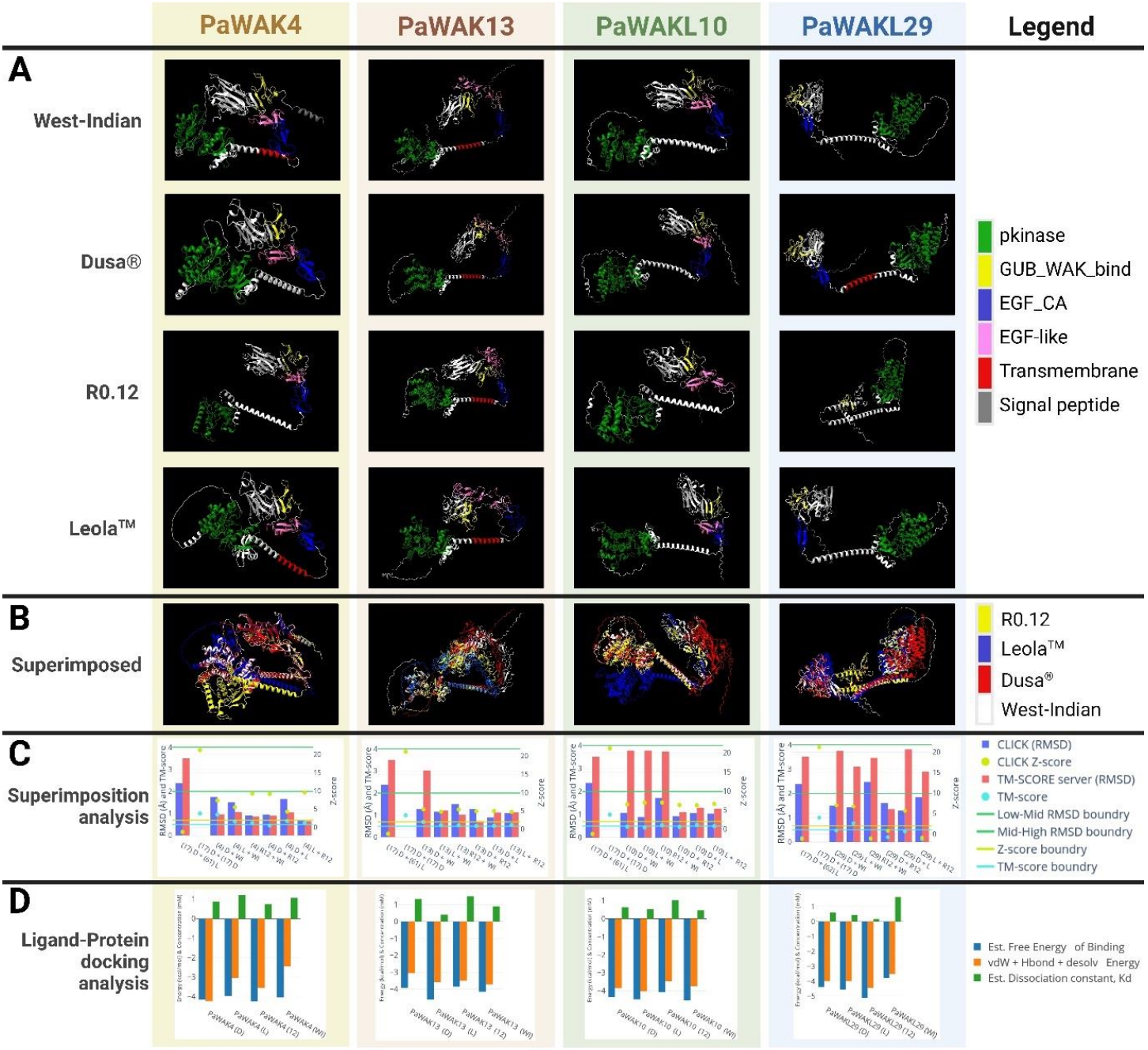


Figure 6: Comparison of the protein structures of PaWAK4, PaWAK13, PaWAKL10, and PaWAKL29 across four rootstocks, and their affinity to bind oligogalacturonides. (**A**) Visualization of the 3D structure of PaWAK4, PaWAK13, PaWAKL10 and PaWAKL29 (**B**) The visual superimposition of the four predicted protein structures from four different rootstock genomes. (**C**) 3D protein root-mean-square deviation (RMSD) scores with associated Z-scores from CLICK server and RMSD scores with associated TM-Scores from TM-score server to assess the similarity of PaWAK/WAKL proteins across four different rootstocks. Boundary lines for RMSD values (green line at 2 Å and 4 Å), Z-score (yellow line at 2 Å), and TM-score (light blue line at 0.5 Å) indicate critical thresholds. Gene names are summarised as (17) for

PaWAKL17, (4) for PaWAK4, (13) for PaWAK13, (18) for PaWAKL18 and (29) for PaWAKL29 in the x-axis. Different PaWAKLs from two genomes '(17) D + (61) L' and the same PaWAKL from the same genome '(17) D + (17) D' were included to act as a negative and positive control. (**D**) Predicted ligand-protein binding of PaWAK/WAKLs and a pectin monomer (oligogalacturonide, PubChem accession: 441476). Energies involved with the precited binding are shown in blue and orange as negative values while the estimated dissociation concentration (K_d) is shown in green as positive values. Rootstocks: WI-West Indian, D-Dusa®, 12-R0.12 and L-Leola™. Graphs visualized with Plotly (https://chart-studio.plotly.com), 3D models visualized with ChimeraX and image created with BioRender.com.

Discussion

This study aimed to characterize the *PaWAK/WAKL* gene family and investigate a subset of its members for their potential involvement in a successful defence response against *P. cinnamomi*. The identification of 14 *PaWAKs* and 62 *PaWAKLs*, and subsequent characterisation, was performed across four avocado genomes. Among these, *PaWAK4, PaWAK13, PaWAKL10, PaWAKL29*, and *PaWAKL31* showed significant upregulation in the partially-resistant rootstock, Dusa®, at three or four time points during infection. The protein products of these candidate genes displayed variation in 3D structure and predicted ability to bind OGs across the rootstocks. Additionally, the two partially-resistant rootstocks, Dusa® and Leola™, exhibited predicted differences, suggesting that their defence responses differ.

In silico characterisation of the *PaWAK/WAKL* gene family

The repertoire of 14 *PaWAK* and 62 *PaWAKLs* identified in avocado is similar to that identified in *Brassica rapa* but represents more than *Arabidopsis* and fewer *WAK*s than apple and pea (Verica and He, 2002, Zuo *et al.*, 2019, Zhang *et al.*, 2020a, Li *et al.*, 2023). The classification definitions used in this study, based on protein domains, are not uniform across recent literature but represent an amalgamation to ensure a comprehensive protocol as described in the materials and methods section (Zuo *et al.*, 2019, Zhang *et al.*, 2020a, Liu *et al.*, 2021, Li *et al.*, 2022, Sipahi *et al.*, 2022, Xia *et al.*, 2022, Yu *et al.*, 2022, Kong *et al.*, 2023, Li *et al.*, 2023, Wang *et al.*, 2023, Yan *et al.*, 2023, Zhong *et al.*, 2023). *PaWAKL5* is an outlier to this protocol – according to an NCBI CDD search, it would be classified as a WAK, but according to the Pfam and SMART database, it would be classified as a WAKL. For this study, PaWAKL5 was classified

as a WAKL, but this classification can be re-evaluated in future as protein domain databases and their predictions improve. There were differences in gene family composition across the four avocado genomes, with 4, 2 and 7 *PaWAKLs* not identified in Dusa®, R0.12 and Leola™, respectively (**Table S3**)**.** Molecular validation of these discrepancies should be done in the future to determine if they are true differences or due to annotation errors.

The chromosomal clustering of genes within the *WAK*/*WAKL* gene family might indicate tandem duplication which produced identical sequences adjacent to the original sequence. These events allow for the expansion of the gene family within species (Kong *et al.*, 2007, Lallemand *et al.*, 2020). Tandemly duplicated pairs of *WAK*/*WAKLs* have been predicted in bread wheat, walnut, cotton, cannabis, and potato (Dou *et al.*, 2021, Zhang *et al.*, 2021, Li *et al.*, 2022, Sipahi *et al.*, 2022, Xia *et al.*, 2022, Yu *et al.*, 2022). In avocado, the only duplicated pair predicted to be under positive selection was *PaWAKL18* and *PaWAKL19*. Interestingly, *PaWAKL18* is also a tandemly duplicated pair with *PaWAKL17*, but this pair was predicted to be under purifying selection (**Figure 1B**). A hypothesis that could explain the positive selection of *PaWAKL18* is that it allows for neofunctionalization, where a new or different function arises for the duplicated sequence (Roulin *et al.*, 2013). In contrast, purifying selection allows for subfunctionalization, the division of the parent gene's function. Thus, *PaWAKL19* could have been tandemly duplicated from *PaWAKL18* and is undergoing selection to develop a novel function. Phylogenetic analysis showed that most of the tandemly duplicated pairs grouped together, providing further evidence of their close relationships (**Figure 4**). The formation of mostly species-specific clades in the phylogenetic analysis indicated that this gene family has expanded independently in different plant species, similar to observations made previously for cotton, barley and tomato (Sun *et al.*, 2020, Tripathi *et al.*, 2021, Zhang *et al.*, 2021).

The PaWAK/WAKLs were predicted to localise predominantly to the plasma membrane, with 24 proteins (31.2%) localised there (**Figure S1**). However, the remaining proteins were localised across seven other cellular locations, similar to predictions for the *Cannabis sativa* WAK/WAKLs (Sipahi *et al.*, 2022). Although all PaWAKs contained a TM domain and signal peptide, not all of them were predicted to be localised at the plasma membrane. The descriptions were based on the West-Indian protein predictions but some discrepancies were seen between the rootstocks. This suggested that the PaWAK/WAKLs perform their functions

throughout the cell and are not restricted to the plasma membrane, potentially embedding in other membranes. This could allow PaWAK/WAKLs to detect OGs after cell wall damage, where OGs could enter the cell along with the pathogen. PaWAK/WAKLs embedded in other organelle membranes could activate pathways specific to those organelles. For example, PaWAK/WAKLs localizing to the mitochondria could influence ROS production, a downstream effect proposed for GhWAK7A-mediated chitin signalling against fungal pathogens in cotton (Wang *et al.*, 2020). The conserved motifs were mostly localized to the C-terminal of the PaWAK/WAKLs, within the kinase domain, suggesting that the kinase domain has the largest conservation across the proteins, similar to observations in barley, cannabis, cotton, pea, potato, sesame, and tomato (Kurt *et al.*, 2020, Tripathi *et al.*, 2021, Zhang *et al.*, 2021, Sipahi *et al.*, 2022, Yu *et al.*, 2022, Li *et al.*, 2023, Yan *et al.*, 2023). The kinase domain performs the fundamental function of phosphorylating proteins, which could explain the minimal variation in these regions. In contrast, the N-terminal showed more variation and contained the GWB and EGF domains. This variation in the N-terminal might allow PaWAK/WAKLs to have specialized cellular roles, enabling a more targeted response to environmental or developmental changes.

*PaWAK/WAKL*s implicated in avocado defence against *P. cinnamomi*

There were 31 differentially expressed *PaWAK/WAKLs* in Dusa® and 23 in R0.12 at one or more time points, with a total of 35 unique genes, indicating that this family is responsive to *P. cinnamomi* infection (**Figure 3**). *PaWAK4*, *PaWAK13*, *PaWAKL10*, *PaWAKL29*, and *PaWAKL31* showed significant upregulation across three or four time points in the partially-resistant rootstock Dusa®. In contrast, upregulation of *PaWAK13*, *PaWAKL29,* and *PaWAKL31* only occurred at 6 hpi in the susceptible rootstock R0.12. These genes were not clustered together following phylogenetic analysis, except for *PaWAK4* and *PaWAKL10*, which showed a very close relationship and were predicted to be a tandemly duplicated pair with a K_a/K_s ratio < 1 (**Figures 1 & 4**). It is possible that *PaWAKL10* was tandemly duplicated from *PaWAK4* and has now undergone subfunctionalization, with both acting together during *P. cinnamomi* infection. The main difference in the expression patterns between the two rootstocks was the duration of differential expression. All *PaWAK/WAKLs* that showed significant up- or down-regulation at 6 hpi in R0.12 subsequently returned to non-significant expression by 12 hpi. In contrast, most differentially expressed *PaWAK/WAKLs* in Dusa® at 6 hpi maintained their

significant up- or down-regulation across multiple time points. This prolonged differential expression in Dusa® could contribute to its successful defence response compared to R0.12.

A phylogenetic analysis was conducted using the PaWAK/WAKLs, AtWAK/WAKLs and dicotyledonous plant WAK/WAKLs, previously implicated in defence, to identify relationships between the proteins and to infer function (**Figure 4**). Most defence-related sequences clustered with the AtWAK/WAKLs, showing minimal dispersion with the PaWAK/WAKLs. Notably, PaWAKL40, PaWAKL43, PaWAKL60, PaWAKL61, and PaWAKL62 were clustered with two cotton GhWAKs, four AtWAKLs and a *Medicago truncatula Mt*WAK. However, all corresponding genes, except PaWAKL40, were significantly downregulated at one or more time points in Dusa®, suggesting they do not contribute to a successful defence response. PaWAK13 exhibited a moderate relationship with cotton GhWAK5A, GhWAK7A and a tomato SlWAK1. GhWAK7A is involved in cotton defence against two fungal pathogens, a hemibiotroph (*Verticillium dahlia*) and a nectrotroph (*Fusarium oxysporum*), through chitin signalling pathways. GhWAK5A showed upregulation seven days post *V. dahliae* inoculation, although the downstream pathway remains unknown (Wang *et al.*, 2020). Despite the low protein percent identity and structural similarity between PaWAK13, GhWAK5A and GhWAK7A (44.87 and 42.2%, respectively, with RDSM values > 3.6 and TM-score < 0.18), it shows the closest phylogenetic relationship in this subset and so there could be functional similarities, as both are also associated with defence responses against hemibiotrophic pathogens. With the limited relationships between the upregulated PaWAK/WAKLs proteins and the defence-related sequences from the other species, the functional inference here is minimal.

The *cis*-acting element composition in the promoter regions of upregulated genes was assessed and compared to predict which pathways may have contributed to the observed differential expression (**Figure 5**). Although expression data for the *PaWAK/WAKLs* in Leola™ are unavailable, the *cis*-acting elements can still provide evidence to suggest which pathways might be involved in their regulation. Salicylic acid (SA), jasmonic acid (JA), ABA, and auxin (phytohormones) were previously linked to the successful defence response of avocado against *P. cinnamomi* (Van den Berg *et al.*, 2018). It would be informative to determine if the *PaWAK/WAKLs* have elements in their promoter regions responsive to these phytohormones to infer which pathways might be involved in their expression. The increased number of TCA-

elements (SA-responsive element) in the Dusa® *PaWAK13* promoter compared to the R0.12 promoter could partly explain the differences in expression observed between these two rootstocks. The CGTCA- and TGACG-motifs (the positive and negative promoter strand), are responsive to methyl jasmonate (MeJA) and were more abundant in the Dusa® *PaWAK13* promoter compared to the R0.12 promoter. This suggests that the Dusa® *PaWAK13* promoter might be more responsive to MeJA or JA than the R0.12 promoter (**Figure 5**). The ABA-responsive element, ABRE, was more abundant in *PaWAK4* compared to other upregulated gene promoters, indicating that it might be more responsive to ABA than the other genes. The Leola™ promoter had the most ABRE elements, suggesting it could be the most responsive to ABA. However, more ABRE elements were found in the R0.12 *PaWAK4* promoter when compared to Dusa®, suggesting this pathway may not be responsible for the increased expression of *PaWAK4* in Dusa® during *P. cinnamomi* infection.

Another important set of *cis*-acting elements to consider are the stress-responsive elements, namely CCAAT-box, STRE and TC-rich repeats. Few TC-rich repeat elements were present in promoter regions of the genes upregulated across three or more time points in Dusa®. However, the Leola™ *PaWAKL29* promoter had TC-rich repeat elements, unlike the R0.12 and Dusa® promoters. The CCAAT-box was present in Leloa™ *PaWAKL10* and *PaWAKL29* but absent in the promoters of these genes in R0.12, suggesting that they may be more responsive to stress in Leola™ compared to R0.12, potentially contributing to its susceptibility. STRE was more abundant in the Dusa® and Leola™ *PaWAK4* promoters compared to R0.12, indicating they may be more stress-responsive in the partially-resistant rootstocks. This could contribute to the differential expression observed in Dusa® during the biotic stress response against *P. cinnamomi* infection and illustrates the differences in promoter composition between these two rootstocks, which could contribute to different levels of involvement in different cellular pathways.

Candidate defence PaWAK/WAKL analyses show structural and ligand-binding differences

The differences in expression and promoter *cis*-acting elements may not be the only factors contributing to varying defence efficiencies between rootstocks. Protein differences could also play a significant role. To explore this, the protein sequences of PaWAK4, PaWAK13, PaWAKL10, and PaWAKL29 were assessed to predict differences in protein structure and their

ability to bind OGs. It is known that *P. cinnamomi* contains polygalacturonase (PG) coding genes, these proteins are involved in pectin cleavage, and three *PG*s show upregulation during infection of avocado, as observed in the RNA-sequencing data used in this study (Backer *et al.*, 2022, Miyambo *et al.*, 2022). This suggests that OGs are produced during infection, making it crucial to identify the proteins responsible for recognizing them.

When visualized, these proteins showed notable discrepancies between the rootstocks (**Figure 6A-C**). The comparison of PaWAK4 in Dusa® and R0.12, PaWAKL10 in Dusa® and R0.12 and PaWAKL10 in Leola™ and R0.12 showed TM-scores < 0.5. This suggested that these proteins have high structural similarity (more than expected by chance due to RMSD score) but may not lie in the same fold, as their 2D structures (α-helices and β-sheets) are in different spatial orientations. PaWAKL10 in the Leola™ and West Indian comparison highlighted a contradiction that may arise from the prediction methods used. The Z-score, which measures the statistical significance of protein alignment and the aligned residues, and the TM-score, which assesses structural similarity based on C-alpha atoms in the aligned residues, yielded different interpretations (Zhang and Skolnick, 2004, Nguyen *et al.*, 2011). Therefore, a high Z-score and low TM-score implied that the alignment was statistically significant, but that the structures do not share the same fold, potentially leading to functional differences. The results indicated that PaWAKL29 in R0.12 differed from its equivalent in West-Indian and Leola™ rootstocks. This provides quantitative evidence to suggest that the same protein can differ structurally in different rootstocks, which can lead to downstream effects. A more in-depth protein structure analysis would be required for stronger conclusions.

Based on their negative binding energies in the ligand-binding analysis (**Figure 6D**), PaWAK4, PaWAK13, PaWAKL10, and PaWAKL29 were identified as candidate defence-related proteins, expected to recognize and bind OGs and influence downstream defence pathways. Differences were observed among the proteins from different rootstocks, with the Dusa® versions of PaWAK13 and PaWAKL10 predicted to bind OGs more efficiently than their R0.12 counterparts. This finding, combined with transcriptomic analyses, provided multi-omic evidence for the involvement of *PaWAK13* and *PaWAKL10* in the successful defence response against *P. cinnamomi*. Additionally, the Leola™ PaWAK13 and PaWAKL10 were predicted to have a higher affinity for OGs than those from both Dusa® and R0.12, implying that these proteins may respond more rapidly to pathogen penetration, potentially initiating the

defence response earlier. This analysis was conducted on WAK/WAKL monomers however protein receptors containing EGF-domains can form homo- and heterodimers and this should be investigated in the PaWAK/WAKL context in the future (Schlessinger, 2002).

Conclusion

This study is the first comprehensive *in silico* characterization of the complete *PaWAK/WAKL* gene family in avocado, highlighting the involvement of *PaWAK4*, *PaWAK13*, *PaWAKL10, PaWAKL29*, and *PaWAKL31* in the defence response against *P. cinnamomi* based on their upregulation across three or more timepoints in the partially-resistant rootstock Dusa® as compared to its susceptible counterpart, R0.12. Promoter analyses indicated that the *PaWAK/WAKLs* that were upregulated at multiple time points in Dusa® were associated with the complex regulatory networks in avocado defence. The structural modelling and binding affinity analyses revealed significant differences among the PaWAK/WAKL proteins across different avocado rootstocks of varying resistance to *P. cinnamomi*. These insights suggested that the continued upregulation of *PaWAK/WAKLs* associated with phytohormone- and stress defence pathways, and the efficient recognition of OGs by these proteins may enhance the early defence response in a partially-resistant rootstock. Future research should focus on validating these candidate gene roles through functional assays and exploring their application in a molecular screening tool to aid in the selection of resistant avocado rootstocks for commercial use.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

AH: Conceptualisation, formal analysis, investigation, project administration, visualization, writing – original draft. NvdB and VS: Conceptualisation, Funding acquisition, resources, supervision, writing – review & editing – All authors consent for the work to be published

Funding

Funding was provided by the Hans Merensky Legacy Foundation and the National Research Foundation (NRF).

Data Availability Statement

The authors confirm that the data supporting the findings of this study are available within the article and its supplementary materials.

Acknowledgments

We would like to acknowledge the Hans Merensky Legacy Foundation for funding this project, the Hans Merensky Chair in Avocado Research for their support, the Avocado Genome Consortium for providing genome data and Alicia Fick for her guidance and mentoring.

Supplementary Material

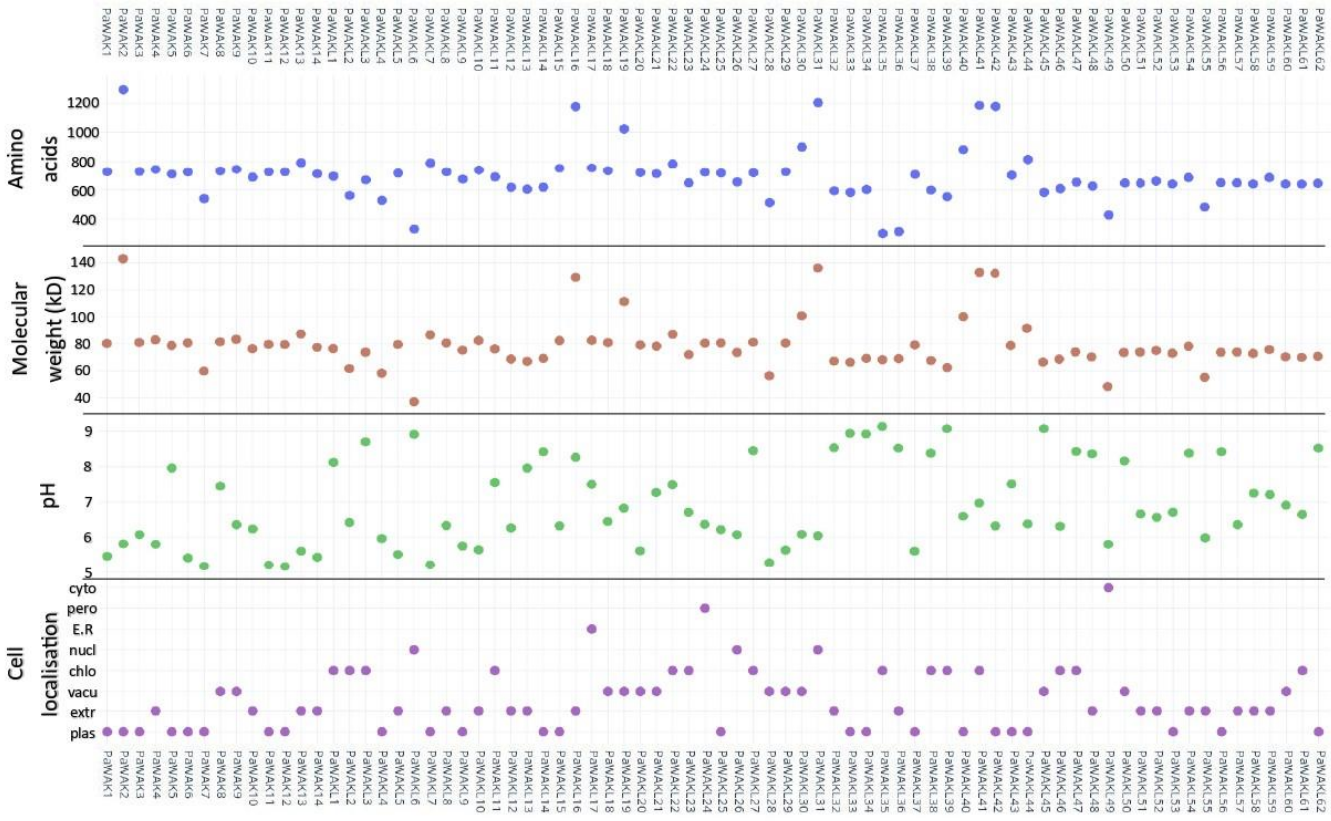


Figure S1: Scatter plots indicating protein length (amino acids), molecular weight (kD), isoelectric point (pH) and the predicted subcellular localisation of the PaWAK/WAKL proteins. Plas-Plasma membrane; Extr-Extracellular space; Vacu-Vacuole; Chlo-chloroplast; E.R.-Endoplasmic reticulum, Cyto- cytoplasm; Pero-peroxisome

Table S1: Protein sequences of defence-related WAK/WAKLs used in the phylogenetic analysis

Table S2: *PaWAK/WAKL* tandemly duplicated gene pairs and associated K_a/K_s ratios

Table S3: *PaWAK/WAKL* gene IDs and sequence analysis results for naming and classification

Table S4: PaWAK/WAKL protein properties

Table S5: Expression values of *PaWAK/WAKLs* during *P. cinnamomi* infection, rounded to two decimal places

Table S6: *Cis*-acting element compositions for all *PaWAK/WAKLs* from PlantRegMap and the phytohormone responsive elements from PlantCARE with functional descriptions where they were available

Table S7: Superimposition scores for pairs of PaWAK4, PaWAK13, PaWAKL29 and PaWAKL29 3D protein structures from CLICK and TM-score server

Table S8: PaWAK/WAKL-oligogalacturonide binding predictions with interacting residues

References

Abeel, T., Van Parys, T., Saeys, Y., Galagan, J. & Van De Peer, Y. (2012). Genomeview: A next-generation genome browser. *Nucleic Acids Res,* 40**,** e12. doi:10.1093/nar/gkr995

Backer, R., Engelbrecht, J. & Van Den Berg, N. (2022). Differing responses to *phytophthora cinnamomi* infection in susceptible and partially resistant *persea americana* (mill.) rootstocks: A case for the role of receptor-like kinases and apoplastic proteases. *Frontiers in Plant Science,* 13**,** 928176.

Bikadi, Z. & Hazai, E. (2009). Application of the pm6 semi-empirical method to modeling proteins enhances docking accuracy of autodock. *Journal of Cheminformatics,* 1**,** 15. doi:10.1186/1758-2946-1-15

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). Tbtools: An integrative toolkit developed for interactive analyses of big biological data. *Mol Plant,* 13**,** 1194-1202. doi:10.1016/j.molp.2020.06.009

Delteil, A., Gobbato, E., Cayrol, B., Estevan, J., Michel-Romiti, C., Dievart, A., et al. (2016). Several wall-associated kinases participate positively and negatively in basal defense against rice blast fungus. *BMC Plant Biol,* 16**,** 17. doi:10.1186/s12870-016-0711-x

Dou, L., Li, Z., Shen, Q., Shi, H., Li, H., Wang, W., et al. (2021). Genome-wide characterization of the *wak* gene family and expression analysis under plant hormone treatment in cotton. *BMC genomics,* 22**,** 1-17. doi:10.1186/s12864-021-07378-8

Finn, R. D., Clements, J. & Eddy, S. R. (2011). Hmmer web server: Interactive sequence similarity searching. *Nucleic Acids Res,* 39**,** W29-37. doi:10.1093/nar/gkr367

Hardham, A. R. (2005). *Phytophthora cinnamomi*. *Molecular Plant Pathology,* 6**,** 589-604. doi:10.1111/J.1364-3703.2005.00308.X

Hardham, A. R. & Blackman, L. M. (2018). *Phytophthora cinnamomi*. *Molecular Plant Pathology,* 19**,** 260-285. doi:10.1111/mpp.12568

Kohorn, B. & Kohorn, S. (2012). The cell wall-associated kinases, waks, as pectin receptors. *Frontiers in Plant Science,* 3. doi:10.3389/fpls.2012.00088

Kohorn, B. D. (2015). Cell wall-associated kinases and pectin perception. *Journal of Experimental Botany,* 67**,** 489-494. doi:10.1093/jxb/erv467

Kong, H., Landherr, L. L., Frohlich, M. W., Leebens-Mack, J., Ma, H. & Depamphilis, C. W. (2007). Patterns of gene duplication in the plant *skp1* gene family in angiosperms: Evidence for multiple mechanisms of rapid gene birth. *The Plant Journal,* 50**,** 873-885. doi:10.1111/j.1365-313X.2007.03097.x

Kong, W., Shi, J., Yang, B., Yu, S., Zhao, P., Guo, Z., et al. (2023). Genome-wide analysis of the *wall-associated kinase* (*wak*) genes in *medicago truncatula* and functional characterization of *mtwak24* in response to pathogen infection. *Plants (Basel),* 12. doi:10.3390/plants12091849

Kurt, F., Kurt, B. & Filiz, E. (2020). Wall associated kinases (*waks*) gene family in tomato (*solanum lycopersicum*): Insights into plant immunity. *Gene Reports,* 21**,** 100828. doi:10.1016/j.genrep.2020.100828

Lallemand, T., Leduc, M., Landès, C., Rizzon, C. & Lerat, E. (2020). An overview of duplicated gene detection methods: Why the duplication mechanism has to be accounted for in their choice. *Genes (Basel),* 11. doi:10.3390/genes11091046

Larkan, N. J., Ma, L., Haddadi, P., Buchwaldt, M., Parkin, I. a. P., Djavaheri, M., et al. (2020). The *brassica napus wall-associated kinase-like* (*wakl*) gene *rlm9* provides race-specific blackleg resistance. *The Plant Journal,* 104**,** 892-900. doi:10.1111/tpj.14966

Lescot, M., Déhais, P., Thijs, G., Marchal, K., Moreau, Y., Van De Peer, Y., et al. (2002). Plantcare, a database of plant *cis*-acting regulatory elements and a portal to tools for *in silico* analysis of promoter sequences. *Nucleic Acids Research,* 30**,** 325-327. doi:10.1093/nar/30.1.325

Li, M., Ma, J., Liu, H., Ou, M., Ye, H. & Zhao, P. (2022). Identification and characterization of *wall-associated kinase* (*wak*) and *wak-like* (*wakl*) gene family in *juglans regia* and its wild related species *juglans mandshurica*. *Genes,* 13**,** 134. doi:10.3390/genes13010134

Li, X., Ou, M., Li, L., Li, Y., Feng, Y., Huang, X., et al. (2023). The *wall-associated kinase* gene family in pea (*pisum sativum*) and its function in response to b deficiency and al toxicity. *Journal of Plant Physiology,* 287**,** 154045. doi:10.1016/j.jplph.2023.154045

Liu, X., Wang, Z., Tian, Y., Zhang, S., Li, D., Dong, W., et al. (2021). Characterization of *wall-associated kinase/wall-associated kinase-like* (*wak/wakl*) family in rose (*rosa chinensis*) reveals the role of *rcwak4* in *botrytis* resistance. *BMC Plant Biol,* 21**,** 1-12. doi:10.1186/s12870-021-03307-9

Miyambo, T. M., Backer, R., Engelbrecht, J., Joubert, F., Van Der Merwe, N. A. & Van Den Berg, N. (2022). The identification and characterization of endopolygalacturonases in

a south african isolate of *phytophthora cinnamomi*. *Microorganisms,* 10**,** 1061. doi:10.3390/ microorganisms10051061

Mohanta, T. K., Khan, A., Hashem, A., Abd_Allah, E. F. & Al-Harrasi, A. (2019). The molecular mass and isoelectric point of plant proteomes. *BMC Genomics,* 20**,** 631. doi:10.1186/s12864-019-5983-8

Nguyen, M. N., Tan, K. P. & Madhusudhan, M. S. (2011). Click—topology-independent comparison of biomolecular 3d structures. *Nucleic Acids Research,* 39**,** W24-W28. doi:10.1093/nar/gkr393

Racine, J. S. (2012). R studio: A platform-independent ide for r and sweave. *Journal of Applied Econometrics,* 27**,** 167-172. doi:10.1002/jae.1278

Ramírez-Gil, J. G., Castañeda-Sánchez, D. A. & Morales-Osorio, J. G. (2017). Production of avocado trees infected with *phytophthora cinnamomi* under different management regimes. *Plant Pathology,* 66**,** 623-632. doi:10.1111/ppa.12620

Ramírez-Sánchez, O., Pérez-Rodríguez, P., Delaye, L. & Tiessen, A. (2016). Plant proteins are smaller because they are encoded by fewer exons than animal proteins. *Genomics Proteomics Bioinformatics,* 14**,** 357-370. doi:10.1016/j.gpb.2016.06.003

Roulin, A., Auer, P. L., Libault, M., Schlueter, J., Farmer, A., May, G., et al. (2013). The fate of duplicated genes in a polyploid plant genome. *The Plant Journal,* 73**,** 143-153. doi:10.1111/tpj.12026

Schlessinger, J. (2002). Ligand-induced, receptor-mediated dimerization and activation of egf receptor. *Cell,* 110**,** 669-672. doi:10.1016/S0092-8674(02)00966-2

Sipahi, H., Whyte, T. D., Ma, G. & Berkowitz, G. (2022). Genome-wide identification and expression analysis of *wall-associated kinase* (*wak*) gene family in *cannabis sativa l. Plants,* 11**,** 2703. doi:10.3390/plants11202703

Stephens, C., Hammond-Kosack, K. E. & Kanyuka, K. (2022). Waksing plant immunity, waning diseases. *Journal of Experimental Botany,* 73**,** 22-37. doi:10.1093/jxb/erab422

Sun, Z., Song, Y., Chen, D., Zang, Y., Zhang, Q., Yi, Y., et al. (2020). Genome-wide identification, classification, characterization, and expression analysis of the *wall-associated kinase* family during fruit development and under wound stress in tomato (*solanum lycopersicum* l.). *Genes,* 11**,** 1186. doi:10.3390/genes11101186

Tian, F., Yang, D.-C., Meng, Y.-Q., Jin, J. & Gao, G. (2019). Plantregmap: Charting functional regulatory maps in plants. *Nucleic Acids Research,* 48**,** D1104-D1113. doi:10.1093/nar/gkz1020

Tripathi, R. K., Aguirre, J. A. & Singh, J. (2021). Genome-wide analysis of *wall associated kinase* (*wak*) gene family in barley. *Genomics,* 113**,** 523-530. doi:10.1016/j.ygeno.2020.09.045

Van Den Berg, N., Mahomed, W., Olivier, N. A., Swart, V. & Crampton, B. G. (2018). Transcriptome analysis of an incompatible *persea americana-phytophthora cinnamomi* interaction reveals the involvement of sa-and ja-pathways in a successful defense response. *PloS one,* 13**,** e0205705. doi:10.1371/journal.pone.0205705

Van Den Berg, N., Swart, V., Backer, R., Fick, A., Wienk, R., Engelbrecht, J., et al. (2021). Advances in understanding defense mechanisms in *persea americana* against *phytophthora cinnamomi*. *Frontiers in Plant Science,* 12. doi:10.3389/fpls.2021.636339

Verica, J. A. & He, Z.-H. (2002). The *cell wall-associated kinase* (*wak*) and *wak-like* kinase gene family. *Plant physiology,* 129**,** 455-459. doi:10.1104/pp.011028

Wang, D., Qin, L., Wu, M., Zou, W., Zang, S., Zhao, Z., et al. (2023). Identification and characterization of *wak* gene family in *saccharum* and the negative roles of *scwak1* under the pathogen stress. *International Journal of Biological Macromolecules,* 224**,** 1-19. doi:10.1016/j.ijbiomac.2022.11.300

Wang, P., Zhou, L., Jamieson, P., Zhang, L., Zhao, Z., Babilonia, K., et al. (2020). The cotton wall-associated kinase ghwak7a mediates responses to fungal wilt pathogens by complexing with the chitin sensory receptors. *The Plant Cell,* 32**,** 3978-4001. doi:10.1105/tpc.19.00950

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). Mcscanx: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research,* 40**,** e49-e49. doi:10.1093/nar/gkr1293

Xia, X., Zhang, X., Zhang, Y., Wang, L., An, Q., Tu, Q., et al. (2022). Characterization of the *wak* gene family reveals genes for fhb resistance in bread wheat (*triticum aestivum* l.). *International Journal of Molecular Sciences,* 23**,** 7157. doi:10.3390/ ijms23137157

Xu, J. & Zhang, Y. (2010). How significant is a protein structure similarity with tm-score = 0.5? *Bioinformatics,* 26**,** 889-895. doi:10.1093/bioinformatics/btq066

Yan, W., Hu, P., Ni, Y., Zhao, H., Liu, X., Cao, H., et al. (2023). Genome-wide characterization of the *wall-associated kinase-like* (*wakl*) family in sesame (*sesamum indicum*) identifies a *siwakl6* gene involved in resistance to *macrophomina phaseolina*. *BMC Plant Biol,* 23**,** 624. doi:10.1186/s12870-023-04658-1

Yu, H., Zhang, W., Kang, Y., Fan, Y., Yang, X., Shi, M., et al. (2022). Genome-wide identification and expression analysis of *wall-associated kinase* (*wak*) gene family in potato (*solanum tuberosum* l.). *Plant Biotechnology Reports,* 16**,** 317-331. doi:10.1007/s11816-021-00739-5

Zhang, B., Li, P., Su, T., Li, P., Xin, X., Wang, W., et al. (2020a). Comprehensive analysis of *wall-associated kinase* genes and their expression under abiotic and biotic stress in chinese cabbage (*brassica rapa* ssp. *Pekinensis*). *Journal of Plant Growth Regulation,* 39**,** 72-86. doi:10.1007/s00344-019-09964-3

Zhang, N., Pombo, M. A., Rosli, H. G. & Martin, G. B. (2020b). Tomato wall-associated kinase slwak1 depends on fls2/fls3 to promote apoplastic immune responses to pseudomonas syringae *Plant Physiology,* 183**,** 1869-1882. doi:10.1104/pp.20.00144

Zhang, Y. & Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics,* 57**,** 702-710. doi:10.1002/prot.20264

Zhang, Z., Ma, W., Ren, Z., Wang, X., Zhao, J., Pei, X., et al. (2021). Characterization and expression analysis of *wall-associated kinase* (*wak*) and *wak-like* family in cotton. *International Journal of Biological Macromolecules,* 187**,** 867-879. doi:10.1016/j.ijbiomac.2021.07.163

Zhong, X., Li, J., Yang, L., Wu, X., Xu, H., Hu, T., et al. (2023). Genome-wide identification and expression analysis of *wall-associated kinase* (*wak*) and *wak-like* kinase gene family in response to tomato yellow leaf curl virus infection in *nicotiana benthamiana*. *BMC Plant Biol,* 23**,** 146. doi:10.1186/s12870-023-04112-2

Zuo, C., Liu, Y., Guo, Z., Mao, J., Chu, M. & Chen, B. (2019). Genome-wide annotation and expression responses to biotic stresses of the *wall-associated kinase - receptor-like kinase* (*wak-rlk*) gene family in apple (*malus domestica*). *European Journal of Plant Pathology*, 153**,** 771-785. doi:10.1007/s10658-018-1591-8